

Deep Learning-based Vehicle Classification using an Ensemble of Local Expert and Global Networks

Jong Taek Lee and Yunsu Chung
Electronics and Telecommunications Research Institute (ETRI)
Daegu, South Korea

Abstract

Vehicle classification has been a challenging problem because of pose variations, weather / illumination changes, inter-class similarity and insufficient training dataset. With the help of innovative deep learning algorithms and large scale traffic surveillance dataset, we are able to achieve high performance on vehicle classification. In order to improve performance, we propose an ensemble of global networks and mixture of K local expert networks. It achieved a mean accuracy of 97.92%, a mean precision of 92.98%, a mean recall of 90.24% and a Cohen Kappa score of 96.75% on unseen test dataset from the MIO-TCD classification challenge.

1. Introduction

Visual analysis on traffic surveillance has recently attracted significant attention in the computer vision community. Vehicle classification and detection have been considered as difficult problems due to the variations of object and camera poses, image quality, lighting and weather conditions. More importantly, the lack of large-scale vehicle dataset has limited applicable methods. Recently, a large-scale vehicle dataset, *CompCars* [11], was released for fine-grained categorization and verification. With the help of this large-scale dataset, Yang *et al.* [11] showed that deep convolutional networks can successfully classify hundreds of different car models.

While *CompCars* dataset focused on fine-grained categorization with hundreds of car models, the classification challenge dataset of the MIOvision Traffic Camera Dataset (MIO-TCD) [1] focused on the categorization of 11 traffic surveillance relevant objects as shown in Figure 1. Although the number of categories in the MIO-TCD classification challenge is much smaller than the number of categories in *CompCars* dataset, the MIO-TCD classification is a highly challenging problem because the dataset acquired at different times and periods by thousands of cameras in a

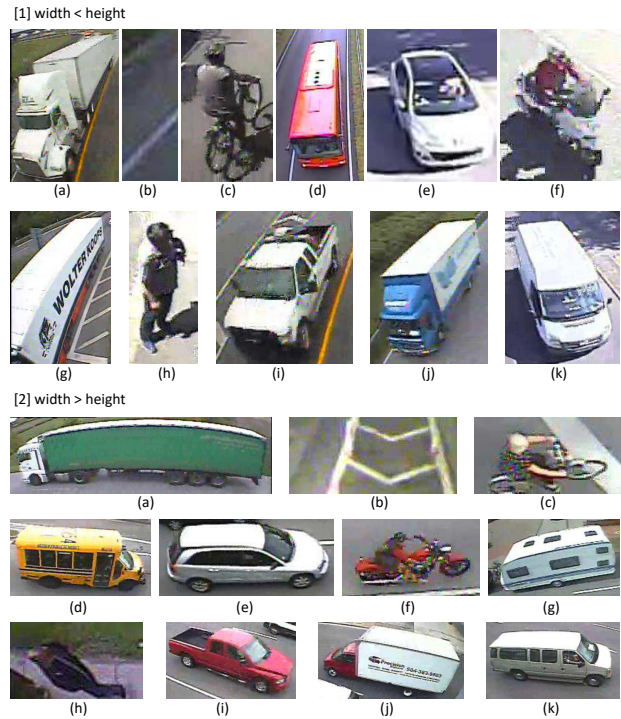


Figure 1: Sample images from 11 categories of MIO-TCD classification dataset: (a) articulated truck, (b) background, (c) bicycle, (d) bus, (e) car, (f) motorcycle, (g) non-motorized vehicle, (h) pedestrian, (i) pickup truck, (j) single unit truck, and (k) work van

wide range of areas. Also, significant inter-class similarity between certain types of vehicles such as articulated truck and single unit truck makes the classification more difficult.

This paper presents an ensemble approach for robust classification on the MIO-TCD challenge. Our system is composed of local expert networks with a gating function [5, 3] and global networks. The local expert and global networks are trained with the particular subsets and entire training set, respectively. An ensemble of these two groups of networks enables our system to reduce an error rate by

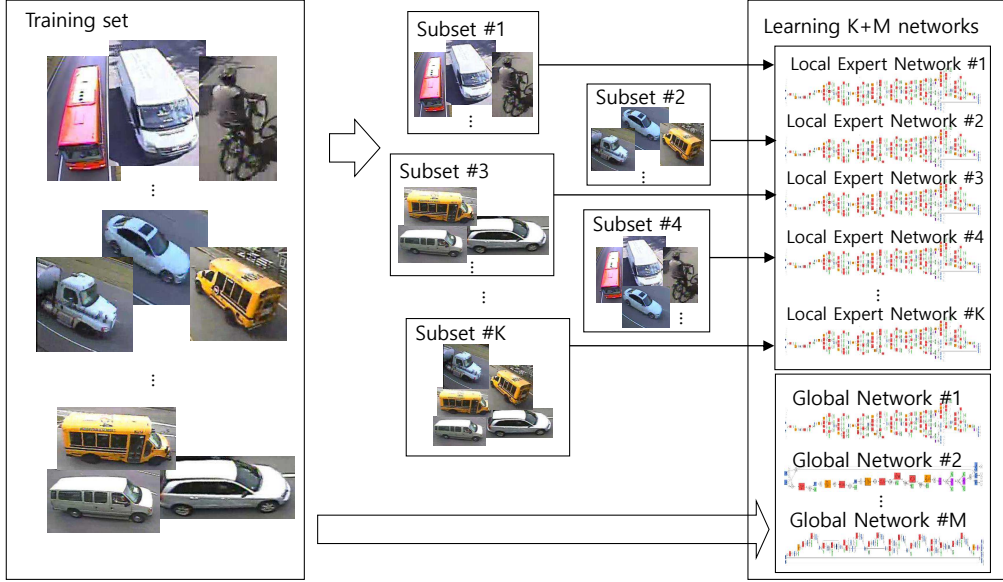


Figure 2: Training phase. Local expert networks are individually trained with each subset of the training data, and global networks are trained with the entire set of training data.

6.5% relative to an ensemble of only global networks. In order to improve individual deep networks of the ensemble, they are trained by applying several classification techniques such as pre-training and multi-crop testing.

The rest of this paper is organized as follows. Section 2 briefly presents prior work related to deep learning based classification networks. We describe the details of our classification framework with training and testing method in Section 3. Section 4 explains the MIO-TCD classification challenge dataset and shows extensive experimental results on the individual networks and their ensemble. Finally, Section 5 concludes the paper.

2. Related work

After Krizhevsky *et al.* [7] presented an outstanding performance in the ImageNet LSVRC (ILSVRC)-2010 contest [8] using deep convolutional neural networks (DCNNs), tremendous deep learning research has been performed to solve classification problems. Szegedy *et al.* [10] proposed a novel deep architecture using inception modules which can increase the depth of networks without boosting the number of parameters. Simonyan and Zisserman [9] showed that 3×3 receptive fields in the first convolutional layer were more effective than 11×11 receptive field with stride 4 [7] or 7×7 with stride 2 [10, 12], and multi-scale training improved performance in the ILSVRC contest. ResNet [3] firstly exceeded the reported human-level performance [8] by using residual learning and parametric rectified linear units. An ensemble prediction of multiple

networks is proven to be effective to reduce error rates in the ILSVRC competition [10, 12, 3].

3. Classification framework

We first generate subsets of training samples by the ratio and size of images, respectively. Local expert DCNNs are individually trained with each subset of the training data, and global DCNNs are trained with the entire set of training data. For testing an image, each DCNN provides a classification results by averaging softmax on multi-crops of the image. The final classification results are calculated by the weighted summation of all DCNN softmax average. The more details of the framework are explained the following subsections.

3.1. Network models

We use three well known deep convolutional neural network structures: AlexNet [8], GoogLeNet [10], and ResNet18 [4]. AlexNet has 8 layers, GoogLeNet has 22 layers, and ResNet18 has 18 layers. In our framework, all of the three structures take 224×224 RGB input images and their last fully convolutional layer has 11 outputs as 11 categories exist in the MIO-TCD Classification challenge.

3.2. Training

The overview of our classification framework for training is shown in Figure 2. In order to train K local expert networks, we generated $K/3$ groups of subsets from the training

set. Each group consists of three subsets, which can be mutually exclusive or partially overlapped. Also, the rule of subset generation is based on the width to height (aspect) ratio of the input images and the size of the input images. All K expert networks have the same GoogLeNet architecture with the weights initiated by the pre-trained ImageNet model, but they are trained by using generated K different subsets.

Unlike K local expert networks, M global networks are trained by using the entire training set. We trained AlexNet, GoogLeNet and ResNet18 models with random initialized weights. In addition to three networks, we trained a GoogLeNet model pre-trained on ImageNet by using three different scales of the images: 224×224 , 240×240 and 256×256 . In total, we trained 18 networks with 12 local expert and 6 global networks. When we have more number of networks in an ensemble, it is hard to see the improvement of classification performance.

Four groups of local expert networks are

3.3. Testing

At test time, an expert network gating function decides which expert networks to turn on. Here, we use a rule based gating function. The conditions of the gating function for K expert networks are same to those of the K subset generation the based on the H/W ratio and size of an input image. The image is then resized to 224×224 and 240×240 for single image and multi-scale testing, respectively. For multi-scale testing, we crop the centered 224×224 , 228×228 , 232×232 , 236×236 and 240×240 from the image and its horizontal flip. Because we want to classify a main (focused) object which are mostly located in the center of images when there are multiple objects in the image, crop is applied only to the centers of the image. The cropped images are resized to 224×224 , and the selected expert networks and global networks classify the images. Finally, a prediction is generated by combining the outputs of the local expert and global networks as shown in Figure 3.

3.4. Implementation details

The image is resized to 240×240 (squash) unless the resize dimension is defined. At every training epoch, a 224×224 crop is randomly sampled from the resized image and its horizontal flip. Our training used stochastic gradient descent with a mini-batch size of 128 and 0.01 initial learning rate. The learning rate is decreased by a factor of 10 every 10 epochs, and learning is stopped after 30 epochs. Our implementation is derived from Caffe library [6] and Nvidia DIGITS [2]. The DIGITS system allows us to easily perform training and testing on multi-GPUs (GeForce GTX TITAN X).

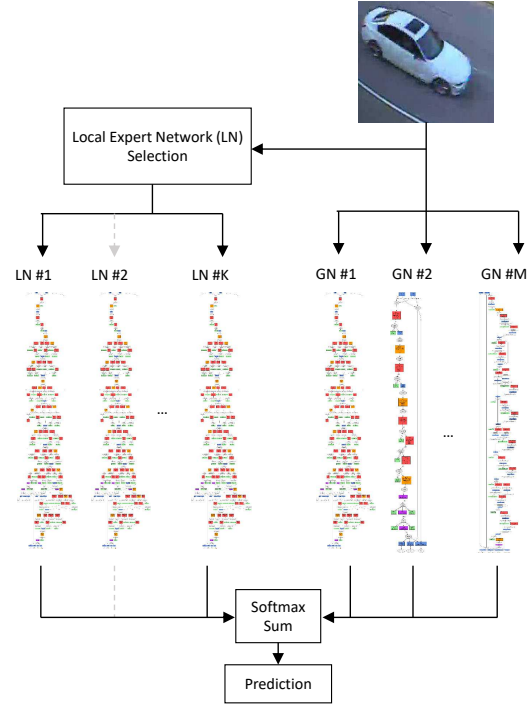



Figure 3: Testing phase. LN and GN stands for local expert network and global network, respectively.

4. Experiments on MIO-TCD classification challenge dataset

4.1. MIO-TCD challenge dataset

The MIO-TCD dataset is a large-scale traffic surveillance challenge for classification and localization. There are 648,959 images for classification and 137,743 images for localization. In this paper, we focus only on the classification challenge dataset. There are 11 traffic surveillance related categories in the classification challenge, including nine types of vehicles, *pedestrian*, and *background*. The size and aspect ratio of an image severely varies in the dataset. One large image can be 18 times larger than a small image, and aspect ratio can be smaller than 0.2 and larger than 14. For example, the aspect ratios of upper images in Figure 1 are smaller than the aspect ratios of lower images in Figure 1. Because the MIO-TCD dataset is collected from real traffic surveillance environments, The counts of *car* and *pickup truck* images are much higher than those of *bicycle*, *motorcycle* and *non-motorized vehicle* images. More detail of image count for category is shown in Figure 4.

Difficult cases. The MIO-TCD classification is challenging due to the interclass similarity and the diversity of pose, lighting and image resolution. More specific difficult cases


Case 1: tow trucks carrying multiple cars








Case 2: two different types of vehicles staying together

Case 3: low resolution images with compression error

Case 4: parts of vehicles not showing

Case 5: very small part of vehicles showing

Case 6: modified vehicles

Case 7: lighting variations at daytime and night

Case 8: fog (left two) and motion blur problem (right two)

Case 9: pedestrian carrying (not riding) a bicycle (left four) and painting on a vehicle (right one)

Table 1: Difficult cases for the MIO-TCD vehicle classification.

are presented in Table 1. In case 1, a huge tow truck carries multiple cars, but only its cargo and carried cars are visible in the images. Their ground truth labels are *non-motorized vehicle*, which makes sense to human, but it can be hard to

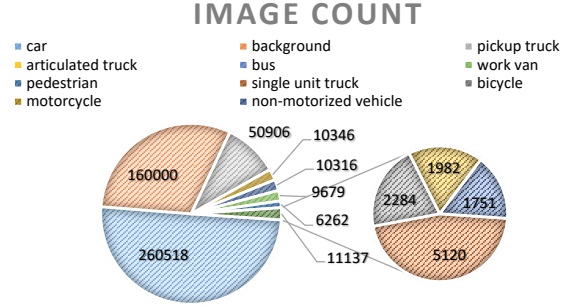


Figure 4: The number of images for each category. Categories are sorted in descending order of image count.

train a classifier with such images. In case 2, two different vehicles stay together and it is hard to choose one vehicle for classification. Their ground truth (GT) labels are *car*. In case 3, the left three low resolution images look like *pedestrian* but their labels are *bicycle*. Also, it is hard to categorize the rightmost image due to compression error. In case 4, the parts of vehicles are missing. Some parts of a vehicle in the two leftmost images are cropped, and the major parts of vehicles in the two center images are occluded while their GT labels are *car*. Case 5 is similar to case 4, but their GT labels are *background* as a very small part of the vehicles is visible. Case 6 presents one of the most confusing cases: vehicle modification. The leftmost three cars look like a wagon(*car*) or a *pickup truck* with a tonneau cover. In case 7 and 8, classifying images is difficult due to color changes or motion blur. In case 9, the GT labels in the four leftmost images are *pedestrian* as people carrying a bicycle instead of riding one.

4.2. Comparative evaluation

We extensively evaluated all M global networks and K local expert networks. Without pre-training, the error rates for single image testing of ResNet18 were 0.3% and 0.5% lower than GoogLeNet and AlexNet, respectively. However, GoogLeNet with weights from the pre-trained ImageNet model reduced the error rate by 19.6% relative to the same network with random weights for single image testing. Figure 5 shows that pre-training helps a network not only learn faster but also converge at a higher accuracy rate. We also compared the kernels of the first convolutional layer of three networks: (a) a pre-trained network, (b) a scratch network trained with whole training data (500K), and (c) a scratch network trained with small training data (1.1K) in Figure 6. A deep neural network can take an advantage when it is trained with a large-scale dataset. The error rates of networks trained with different scales of images became similar when we applied multi-crop testing.

As shown in Table 2, a few of local experts achieved bet-

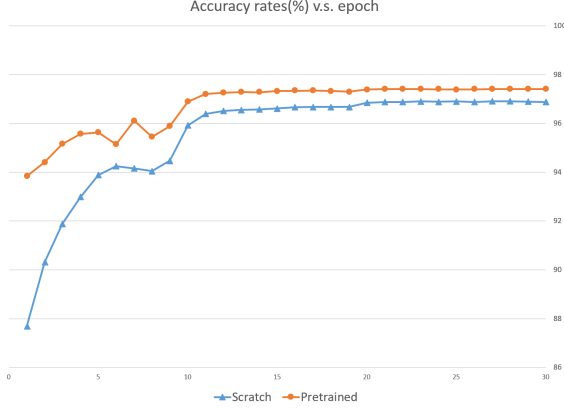


Figure 5: Comparison of network learning with / without pre-training, showing faster convergence and higher accuracy when using pre-trained weights.

ter performance than the global networks. The lowest error rate of expert networks was 1.73% and the biggest error rate was 3.79%. Also, an ensemble of 12 local expert networks achieved 2.15% error rate for single image testing, which is slightly better than an ensemble of 6 global networks. In almost all networks, error rates were reduced when the network tested with multi-crops. However, the reduction became small when a network achieved a very low error rate or we use an ensemble of multiple networks. The error rate of an ensemble of all networks (18) with multi-crop testing was reduced by only 3.1% relative to the one with single image testing. With an ensemble of all trained network by multi-crop testing, we were able to achieve a 2.0% error rate on vehicle classification.

A confusion matrix of the proposed vehicle classification network on the MIO-TCD challenge unseen test set is shown in Table 3. The classification accuracy rates of *car* and *background* are higher than other categories as the samples of these two classes are dominant. *Articulated truck* and *single unit truck*, *work van* and *car*, and *non-motorized vehicle* and *single unit truck* are pairs hard to distinguish. Recall was the lowest on non-motorized vehicle, 68.72%, and the highest on background, 99.66%.

5. Conclusion

In this work, we evaluated various deep convolutional neural networks and their ensemble for large scale traffic surveillance image classification. We demonstrated that multi-crop testing and model training with local expert and global networks is effective with an ensemble of them. Our approach achieved a 98.0% accuracy rate on validation set and a 97.92% accuracy rate on unseen test set in MIO-TCD classification challenge. Future work will explore cluster-

Table 2: Mean error rates of global and local expert networks and its ensemble. IN refers to using pre-trained ImageNet models. ME represents mutually exclusive subsets, and OL indicates overlapping subsets. Error rates in Parentheses of local experts are generated from the entire test set.

Method	Error rates(%)	
	Single	Multi-crop
Global Networks		
AlexNet	3.292	2.887
GoogLeNet	3.091	2.744
ResNet18	2.804	2.531
GoogLeNet_IN_224	2.501	2.448
GoogLeNet_IN_240	2.603	2.421
GoogLeNet_IN_256	2.737	2.500
Local Expert Networks		
GoogLeNet_IN_Ratio_1_ME	3.79(6.50)	3.74(6.59)
GoogLeNet_IN_Ratio_2_ME	2.70(5.66)	2.64(5.61)
GoogLeNet_IN_Ratio_3_ME	1.73(9.30)	1.89(9.2)
GoogLeNet_IN_Ratio_1_OL	3.10(4.01)	3.06(3.97)
GoogLeNet_IN_Ratio_2_OL	2.56(4.84)	2.47(4.75)
GoogLeNet_IN_Ratio_3_OL	2.02(6.03)	1.92(5.84)
GoogLeNet_IN_Size_1_ME	2.80(14.74)	2.72(15.00)
GoogLeNet_IN_Size_2_ME	3.33(7.14)	3.20(6.44)
GoogLeNet_IN_Size_3_ME	2.13(15.39)	2.11(14.42)
GoogLeNet_IN_Size_1_OL	2.99(7.94)	2.84(7.97)
GoogLeNet_IN_Size_2_OL	3.16(4.87)	2.95(4.58)
GoogLeNet_IN_Size_3_OL	2.18(7.43)	2.11(6.71)
Ensemble of		
Global Networks (6)	2.176	2.144
Local Expert Networks (12)	2.151	2.114
All Networks (18)	2.070	2.005

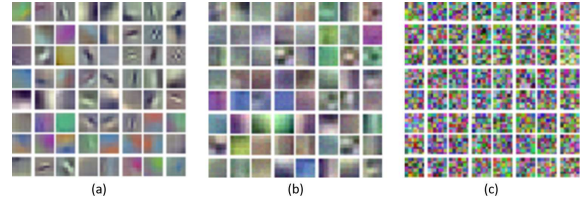


Figure 6: Comparison among kernels of size $7 \times 7 \times 3$ learned by the first convolutional layer of a network with (a) ImageNet pre-trained, (b) a random initial weights trained with 500K images, and (c) a random initial weights trained with 1.1K images.

ing methods for generating K subsets and low resolution image enhancement for tiny image classification.

References

- [1] The miovision traffic camera dataset (mio-tcd). <http://podoce.dinf.usherbrooke.ca/challenge/dataset/>. Accessed: 2017-04-28.

Table 3: A confusion matrix of our ensemble of local expert and global networks on MIO-TCD classification challenge test set.

true	prediction (%)										
	articulated truck	bicycle	bus	car	motorcycle	non-mtr. vhc.	pedestrian	pickup truck	single unit truck	work van	background
articulated truck	93.58	0.00	0.19	0.50	0.00	0.62	0.00	0.15	4.72	0.08	0.15
bicycle	0.00	87.74	0.18	0.70	4.20	0.18	6.65	0.00	0.00	0.00	0.35
bus	0.35	0.00	96.20	0.43	0.00	0.58	0.00	0.12	0.58	1.59	0.16
car	0.01	0.00	0.00	98.89	0.00	0.01	0.00	0.88	0.01	0.15	0.05
motorcycle	0.00	0.61	0.00	5.66	92.12	0.40	0.40	0.81	0.00	0.00	0.00
non-mtr. vhc.	4.79	0.23	0.23	9.13	0.46	68.72	0.46	2.05	10.27	1.60	2.05
pedestrian	0.00	2.68	0.00	1.02	0.83	0.13	94.25	0.00	0.00	0.00	1.09
pickup truck	0.00	0.00	0.03	4.56	0.00	0.06	0.00	95.07	0.18	0.08	0.02
single unit truck	7.73	0.00	0.23	2.11	0.00	0.63	0.00	4.53	82.89	1.56	0.31
work van	0.00	0.00	0.21	14.00	0.00	0.04	0.00	1.57	0.66	83.53	0.00
background	0.01	0.00	0.01	0.28	0.00	0.02	0.02	0.00	0.00	0.00	99.66

- [2] Nvidia interactive deep learning gpu training system (digits). <https://developer.nvidia.com/digits>. Accessed: 2017-04-28.
- [3] Z. Ge, C. McCool, C. Sanderson, and P. Corke. Subset feature learning for fine-grained category classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2015.
- [4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [5] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
- [6] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [11] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3973–3981, 2015.
- [12] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.