Car Classification using Transfer Learning and Siamese Network

Youbin Mo youbinmo@ucsd.edu
Xiaoyin Yang x4yang@eng.ucsd.edu
Xiangyi Dong xid008@ucsd.edu
Xiaochen Liu xiaochen@eng.ucsd.edu

Abstract

Vehicles identification is a critical issue that has been widely discussed and implemented in self-driving car. Correctly identifying the vehicles surrounding the self-driving car is an important criterion for the self-driving control system. In this paper, aiming to improve the car classification accuracy and efficiency by utilizing the existing model, we applied transfer learning by using MobileNet [4], Inception-V3[11] and ResNet-152 [3] pre-trained models to classify the Stanford car classification dataset.[7] Besides, we also attempted to apply Siamese network on it based 2 customized CNNs. By comparing different combined methods, transfer learning largely improves the training speed and the pretrained ResNet has the highest classification accuracy.

1. Introduction

Self-driving car , also known as autonomous car[4], is a kind of vehicle that is able to adjust moving by detecting or sensing its environment with little human control. During the past tens of years, self-driving car has been explored in several points[4]: car classification, distance measurement, automatic moving system, etc. Vehicle identification system is an vital criterion for self-driving system.

Car classification is not only important to self-driving system, but also play a fundamental role on other car-related applications such as object detection, which has applications in many areas of computer vision[12]; electronic toll collection(ETC) system [8], where car classification may correct charges without traffic delay of toll road; traffic surveillance camera [9], which is widely used in the city traffic management(road capacity, traffic density measurement, speed detection, traffic violation detection).

Currently there are SVM models[13] and DNN models [14] on identifying vehicle and non-vehicle images. There still be issues that the identification rate is relatively low and they can only determine whether the object is vehicle or not. However, car classification will be applied on the car model classification or even more in detail, we decide to use transfer learning obtained by GoogleNet and VGGNet, etc. and the other alternative is to apply Siamese networks onto it. By comparing different combined methods, a better performance on car classification and more accurate prediction was proposed in this paper.

2. Related Work

2.1. Classification Problem

Classification problem is one of the most basic essentials in the machine learning area. It belongs to one of the supervised learning problem. And the other one is regression problem. Supervised learning refers to an intelligent task of learning a function that maps an input to an output based on example input-output pairs. The significant difference between supervised learning and unsupervised learning is the former has labeled training data consisting of a set of training examples, while the latter does not have training labels.

For the classification problem, a training dataset is the primary requirement, e.g. images of one specific type, as well as the labels of every images. Then the input feature representation of the learned function is needed to be determined. The accuracy of a learned function strongly depends on which features it chooses during training process. For image input, it is usually to transform the original image into a vector feature space with high dimensions. However, the number of features should be limited while choosing features; otherwise it might stuck in the curse of dimensionality, which also affects the accuracy of prediction. After, one must decide what kind of methods and architecture is going to be implemented. Finally, a test dataset is identified as an evaluation to measure the accuracy of the current classification model.[6]

Recent years, as the improvement of machine learning, there are a lot of algorithms working for basic classification problems. For example, Support Vector Machines, K-Nearest Neighbor algorithm, artificial Neural Networks, etc. Based on Artificial Neural Networks, a lot of more efficient neural network architectures have been invented recently, e.g. Convolutional Neural Networks, Recurrent Neural Networks, and the one with more complicated layers, e.g. ResNet, AlexNet, GoogLeNet, etc. Those more complicated networks are aiming to solve the larger and harder dataset with using complicated and deep-layer net-

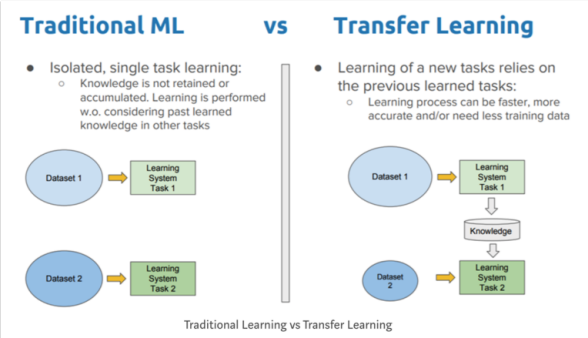

Figure 1. Comparison between traditional Machine Learning and Transfer Learning.

work architecture. However, the more complicated networks are, the longer time it takes to train. In our project, in order to further improve our classification accuracy, we applied ResNet in our car classification problem. ResNet details will be introduced in the next chapter.

2.2. Transfer Learning

Transfer learning is a method that focuses on storing pre-owned knowledge gained in original learning system using original dataset and training process while applying it to a different but related learning system.[10] For example, knowledge gained when learning to recognize flowers could apply when trying to recognize cars. The main advantage of this method is that it can save a lot of time by using pre-trained model and/or initializing a new training process with pre-trained weights. For some complicated network structures, like ResNet mentioned earlier, to train it from scratch usually takes several days even several weeks. However, by applying pre-trained model, it is easily to increase the training speed and shorten the computational cost into several hours. Another benefit of transfer learning is that a lot of pre-trained models are trained on very large dataset like ImageNet, of which the features are very sufficient. Thus, it does not require a large amount of data to train on a new dataset. One can achieve quite ideal accuracy by learning a relatively small new dataset.

2.3. Siamese network

The Siamese network is a special type of neural network architecture with two or more identical subnetwork, whose goal is to find the similarity or comparing the relationship between two comparable things.[5] The main idea is that they will learn important and unique features from the training data which will then further used to compare between the inputs of the respective subnetworks. It learns to differentiate between two inputs and the similarity by differentiating between 2 networks output, instead of one simple model learning to classify its inputs in which contains only one feedforward and one backward learning function. [6] Typical subnetworks commonly used in industry or academics include Convolutionary Neural Network for image data and Recurrent Neural Network for sequential data such as sentences or time series.

A Siamese network consists of two identical neural networks, each taking one of the two input images [11]. The last layers of the two networks are fed to a contrastive loss function (see eq.) , which calculates the similarity between the two images. It is very crucial that not only the architecture of the 2 subnetworks is identical, but the weights are shared among them in order to be called siamese. Therefor, siamese network can learn useful data descriptors that can be further used to compare between the inputs of the respective subnetworks [6].

In binary classification, siamese networks will determine if the inputs are of the same class or not [4]. Like trational classification problem, we can use various loss function during training the network. One of the commonly used loss function is the binary cross-entropy loss which is calculated as follows:

$$L = -y * logp + (1 - y) * log(1 - p)$$

[1], where y is the class label, and p is the predicted value. The network aims to distinguish between similar and dissimilar car images in our case, we will feed one positive and one negative training sample at a time and add up their losses [5]:

$$L = L_+ + L_-$$

However, siamese network typically uses contrasitive loss or triplet loss function [2]. The triplet loss function has the formula as follows:

$$L = \max(d(a, p) - d(a, n) + m, 0)$$

where d is the distance function, a is one training sample, p is any random positive sample and n is any negative sample. m is the pre-defined margin which is used to indicate the separation between the positive and negative samples.

3. Proposed method

3.1. ResNet

ResNet (Residual Neural Network) is a deep residual learning framework that introduced a bypass construction on solving out the degradation problem on deep-layer network.[3] The ResNet used in this project is the original ResNet-152 framework which contains 152 convolution layers that grouped into 34 identity-function-shortcut blocks. As illustrated in Fig.3, two convolution layers in one block learn the input x from higher layer then pass the residual $F(x)$ to lower layer in a feedforward procedure in which causes information loss. Hence, an identity skip-connection is established between the input and the output.

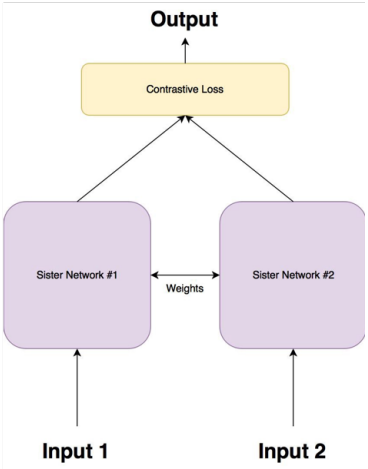

Figure 2. Architecture of a Siamese Network.

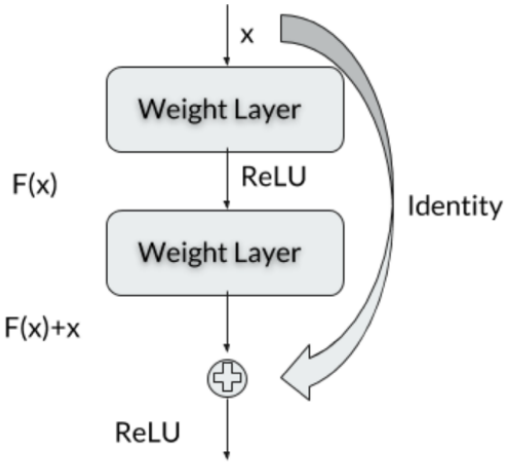

Figure 3. A block of deep residual learning framework.

The deeper layers are able to access information from top layer. In detail, assume x_i is the input of the i-th block and $H(x_i)$ is the learned characteristic. The residual after learning is $F(x_i) = H(x_i) - x_i$. The information transmission function is $x_{i+1} = ReLU(F_i(x_i) + x_i)\ for\ i\ in\ 1, 2, , 34$ where $ReLU(\cdot)$ is an activate function. Therefore, the learning characteristic from layer l to layer L is

$$x_L = x_l + \sum_{i=l}^{L} F(x_i) \tag{1}$$

According to the chain rule, the gradient is

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i)\right) \tag{2}$$

The term $\frac{\partial loss}{\partial x_L}$ is the gradient of loss function to layer L and the 1 in the last term means the information learned by layer L is not zero.

3.2. Inception V3 & MobileNet

Inception architecture (sometimes associated with GoogleNet) is a computational efficiency and fewer parameters machine learning module.[11] With less parameters, 42-layer deep learning network, with similar complexity as VGGNet, can be achieved. The Inception module essentially only computes 1×1 filters, 3×3 filters and 5×5 filters in parallel, but applied bottleneck 11 filters before to reduce the number of parameters. Fig.4 is the Inception Module which has three different configurations of 35×35, 17×17 and 8×8. These Inception Modules only appear in the back of the network, and the front is a normal convolutional layer. Inception V3 uses branching in the branch (8×8 structure) in addition to branching in the Inception Module.

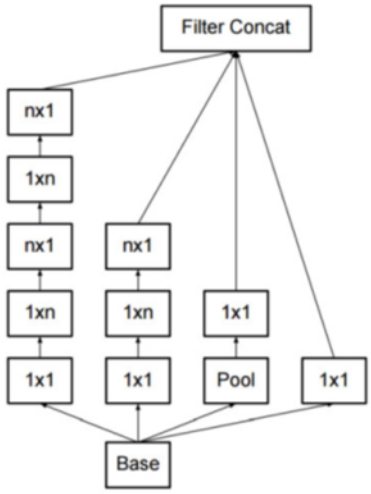

Figure 4. The architecture of Inception-V3 network.

3.3. customized CNN in Siamese network

Our approach follows Hadsell et al [2] and will utilize siamese network to do dimension reduction and image retrievel by computing the Euclidian distance on the output of the shared network [1]. It optimizing the contrasitive loss which is defined as follows:

$$L = \frac{1}{2}(Y)(D)^2 + \frac{1}{2}(1-Y)(\max(0, m-D))^2$$

where $D = \sqrt{(N(x_1 - x_2))^2}$ is the distance between the output of siamese network N on the input x_1 and x_2. The loss function can be spliited in two parts, where:

$$L_{\text{similarity}} = \frac{1}{2}(Y)(D)^2$$

representing the distance of the pair (x_1, x_2) and

$$L_{\text{dissimilarity}} = \frac{1}{2}(1-Y)(\max(0, m-D))^2$$

indicating the penalty of the pair whose distance between them is lower than the pre-defined margin m.

We used five layers of convolutional layer and pooling in our experiment and we also convert full convolutional net to convolutional layer due to the efficient computation on GPU. The model is shown as the Fig. 5 and Fig. 6.

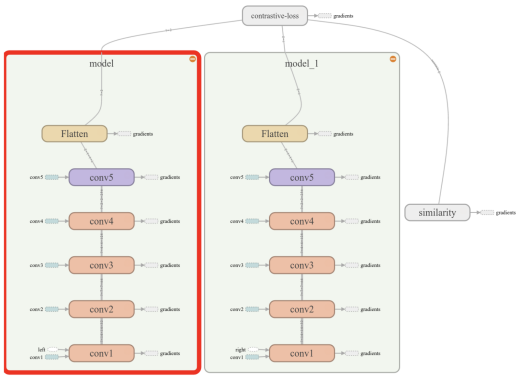

Figure 5. Siamese network used for our car dataset

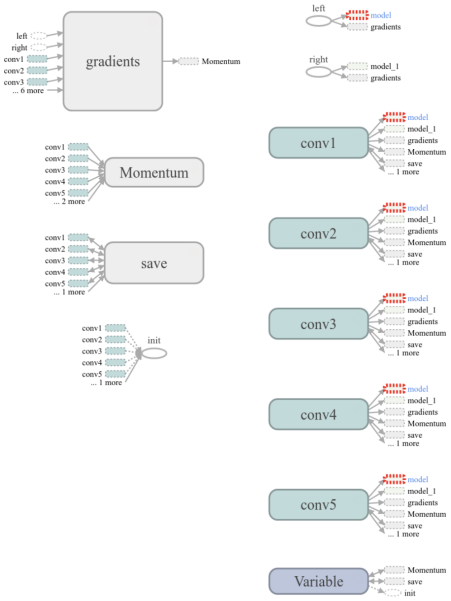

Figure 6. Convolutional layer details used in siamese network

4. Experiments

4.1. Datasets - Stanford Car Dataset[7]

The Cars dataset contains 16,185 images of 196 classes of cars.(see part of dataset in Fig.8) Classes are typically at the level of Make, Model, Year, e.g. 2012 Tesla Model S or 2012 BMW M3 coupe. The data is already split into 8,144 training images and 8,041 testing images, where each class has been split roughly in a 50-50 split. We apply 10% of training dataset as a validation set.

4.2. Data pre-processing

We pre-processed our dataset and cut off useless background in order to have larger fraction of car inside per image. In the dataset website, the bounding box labels are provided for both training and testing datasets. We use the bounding box information to crop the raw images and also applied reflection, rotation and color-filtering onto cropped images. Finally, we have a pre-processed dataset with size of 32,370 (80% for training and 20% for test). (Demonstration in Fig.9)

4.3. Implementation Details

(i) Car Image dataset was downloaded from Stanford database.

(ii) Pre-processing: Cropped the car from each images according to the provided position, and made the mirror images as the supplementary training set.

(iii) Training: Initiated the parameters with the pre-trained models (ResNet, Inception-V3 and MobileNet) then trained the model with pre-processed images for 4000 steps.

(iv) Hardware: CPU Intel-i7 6700K@4.00MHz and GPU NVIDIA GeForce GTX 1080.

(v) Parameters: Train_batch_size = 10, Validation_percentage = 10%, Learning_rate = 0.01.

5. Results

5.1. Training and Testing Error Upon Iteration

Figure.X is the first training and testing result we have. The plot of accuracy indicates that Inception V3 has the lowest training and test accuracy due to its relatively simple architecture. Apparently, ResNet has the highest training and testing accuracy since its architecture is complicated 152-layer pre-trained ResNet model. However, the highest accuracy is only around 70%. (Shown as Fig.7)

According to the first experiment, we already know that ResNet has the highest accuracy. So after applying the cropping, reflection, rotation and color-filtered implementation, we trained the pre-processed new dataset using ResNet. The training accuracy for the new dataset approaches to 90% after 4000 training steps. (Shown as Fig.10)

5.2. Time vs Accuracy

Training speed: 1000 steps/min, on the machine described in Implementation.

Accuracy: reaches 90% above after pre-processing the data

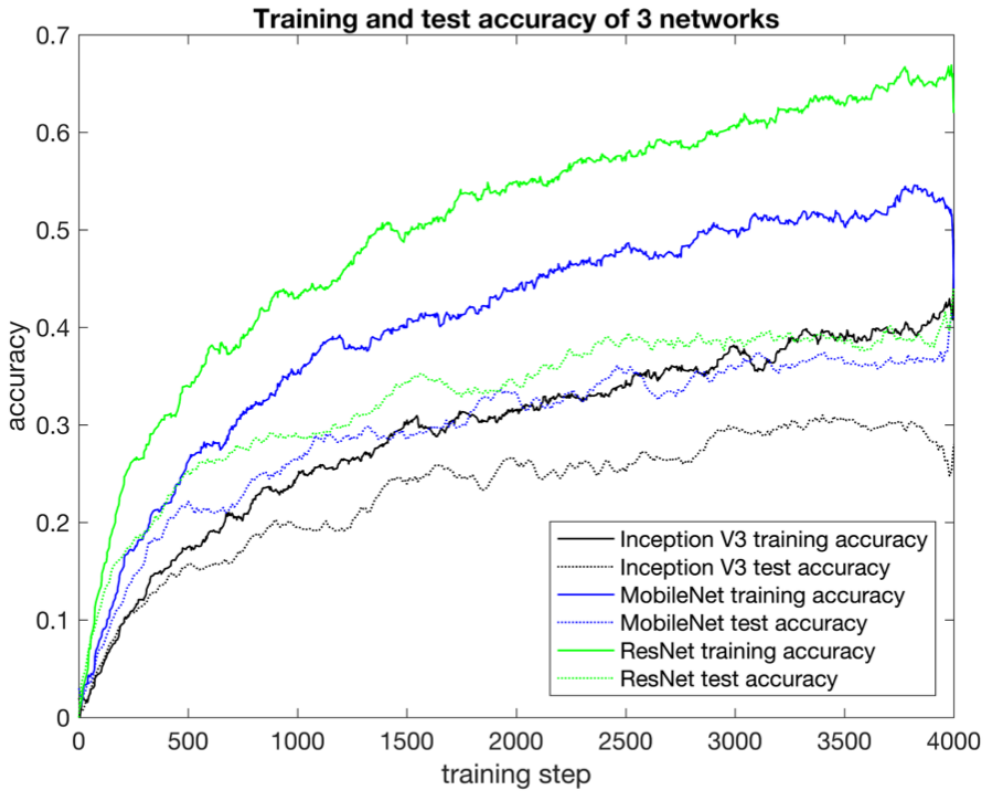

Figure 7. Training and testing accuracy using raw images dataset among MobileNet,Inception and ResNet

Figure 8. Part of images from Stanford Car Dataset

Training step: with more training step(after 20k steps), the final accuracy would stay at 92%.

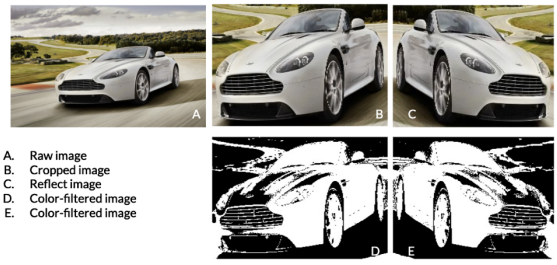

Figure 9. Image pre-processing demonstration. A:raw image; B: cropped image using bounding box; C: rotated image; D and F: color-filtered images

5.3. Pictures of the Best-5 and Worst-5

Shown in Fig.11 and Fig.12, all of the best 5 images have very clear difference between foreground and background, whereas the worst 5 images are harder to distinguish foreground and background. The learning network we developed in this paper has strong ability on identifying the car image with unique color, clear outline and full side-view. Besides, one the worst case contains only the front part of vehicle. Since the training set has few images like that, the probability of correctly classification is low.

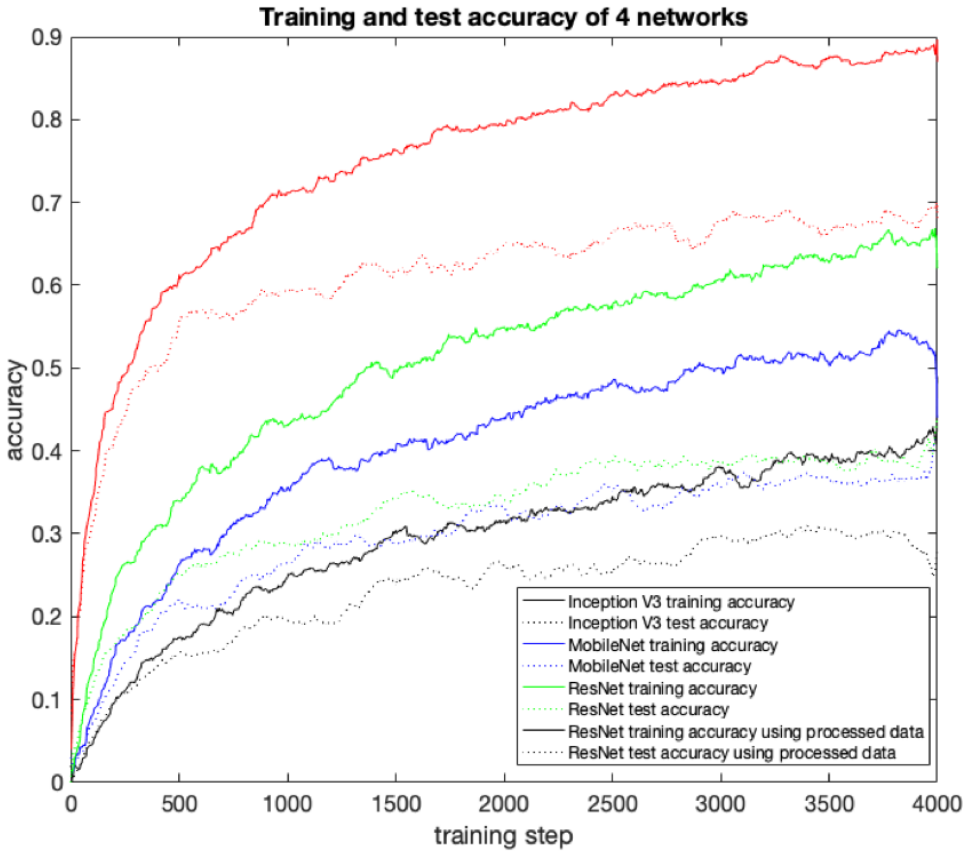

Figure 10. Training and testing accuracy using pre-processed dataset

Figure 11. Best 5 images using ResNet with highest classification accuracy

Figure 12. Worst 5 images using ResNet with lowest classification accuracy

6. Conclusion and Discussion

In this paper, transfer learning method was used to classify the Stanford Car Dataset. The MobileNet, Inception-V3 and ResNet-152 models were applied as pre-trained models, which are all pre-trained using ImageNet dataset. First, we used raw images as dataset, and achieved around 70% accuracy using pre-trained ResNet. Then several image pre-processing methods was implemented and finally pre-trained ResNet model achieved above 90% accuracy. Also the training speed has increased dramatically compared to training from scratch, which can reach 1000 training steps per minute using tensorflow based on GPU NVIDIA GeForce GTX 1080.

For siamese network, siamese network with 2 identical CNN subnetworks was first considered. However, the training result is not very good since the architecture of CNN is too simple. Therefore, building 2 identical ResNet subnet-

works is to be considered next. It is expected to see better result using ResNet subnetworks. Github Repo Link: `https://github.com/Xiaoyin96/ECE271B`

References

[1] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network, 1994.

[2] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping, 2006.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2016.

[4] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[5] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition, 2015.

[6] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[7] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization, 2013.

[8] J. Y. Ng and Y. H. Tay. Image-based vehicle classification system. *arXiv preprint arXiv:1204.2114*, 2012.

[9] K. Robert. Night-time traffic surveillance: A robust framework for multi-vehicle detection, classification and tracking. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6. IEEE, 2009.

[10] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning, 2016.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, 2016.

[12] P. Viola, M. Jones, et al. Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1:511–518, 2001.

[13] M. V. Yeo, X. Li, K. Shen, and E. P. Wilder-Smith. Can svm be used for automatic eeg detection of drowsiness during car driving? *Safety Science*, 47(1):115–124, 2009.

[14] Y. Zhou and N.-M. Cheung. Vehicle classification using transferable deep neural network features. *arXiv preprint arXiv:1601.01145*, 2016.