Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220531

PT module -A Traffic Signal Classification Model Based on Convolutional Neural Networks and Random Forests

KEYU REN[1,*], HEQING PENG[2], JUNWEI WU[3], SHENGTAO YAO[4], JINFENG LI[5], PINGYU LI[6]

[1]School of Software Engineer,Chongqing University of Posts and Telecommunications, Chongqing,400065, China
[2]School of artificial intelligence,Hebei University of Technology, Tianjin, 300131, China
[3]School of Information Science and Technology, Xiamen Univerisity, 43900 Sepang, Selangor Darul Ehsan, Malaysia
[4]Jiangsu Tianyi High School,Wuxi , 214101, China
[5]Huamei Bond International College, Guangzhou, 510520, China
[6]Saint Paul American School, Beijing, 100192, China

*15508037966@163.com

Abstract. Autonomous driving and image recognition are the hot directions of Internet development nowadays. In self-driving cars it is necessary to capture traffic signs in front of the vehicle by cameras. To ensure that the information of image recognition is correct, a set of image classification models with high accuracy should be used to classify the recognized objects in order to determine the next instruction of vehicle operation. There were many achievements in the research work of traffic sign recognition, but there are still some shortcomings. By combining CNN and random forests with PT module (a module that can improve the accuracy of feature extraction), we finally came up with a classification model that can efficiently place the received traffic sign images into a specified category, which we obtained 97% accuracy on the GTSRB dataset, which is much more accurate than traditional neural network methods or regression methods. We have also evaluated it on other datasets and the results obtained are more promising.

Keywords: Deep learning, Convolutional neural networks, Random Forests, Hyperparameter optimization, and image classification

1. Introduction

Autonomous driving technology is divided into 3 core processes in terms of business processes, which are environment-aware localization, decision planning, and execution control. The specific values of these 3 core processes are as follows:

1）Environmental awareness and positioning: mainly through sensor technology and camera, GPS and other technologies to obtain the environmental indicators of the car driving process, and the data will be collected. 2) Decision planning: through the collected data, to make judgments and guid-

DOI: 10.54254/2755-2721/2/20220531

ance on the next behavior of the vehicle. 3) Implementation of decision-making: most of the vehicles are currently using wire control design, how to make decisions through signal commands to control the car's throttle, brake, etc. Related Systems

This paper will focus on studying the algorithm for capturing and classifying traffic signs in environment awareness and being able to match well on our classification model after the camera captures the image. The main techniques used in this are deep learning and random forest.

Deep learning has also become increasingly widespread and its applications in various fields are becoming more and more widespread. neural networks have been used in applications including automation, gaming, and recommendation systems have far surpassed support vector machines and linear regression. At the same time, there have been significant advances in the hardware conditions that support the discipline, with the creation of new GPUs such as the RTX series of graphics cards that support ray tracing, and DLSS systems for optimizing the frame rate of highly detailed image video.

2. Research motivation and guidance

Since the second half of 2018, phenomenal events in the global autonomous driving industry have occurred frequently, and the prologue of commercialization has been kicked off. Autonomous driving can be applied not only to traffic and transportation, but also to logistics and transportation, urban planning, etc. It will produce fundamental innovation. Today, the problem of autonomous driving is mainly focused on behavior prediction, path planning and image segmentation. At this point, we need the car itself to "actively learn" how to drive. The most fundamental of these is the ability for the car to "see" what is in front of it. In Wenhui Li et al. used a branching asymmetric neural network structure and achieved 94% accuracy in GTSRB [1]. 95.7% accuracy was achieved by Ayoub ELLAHYANI et al. using random forest with SVM [2]. Based on their research results, this paper decided to combine CNN with RF and optimize the speed and accuracy of the operation by changing the structure of CNN.

3. Specific method analysis

3.1. Feature extraction

Feature extraction and matching is an important task in many computer vision applications and is widely used in areas such as motion structure, image retrieval, and target detection. Feature extraction and matching consists of three steps: keypoint detection, keypoint feature description and keypoint matching. Main components of feature extraction and matching:

1) detection: identification of points of interest

2) Description: Describes the local appearance around each feature point, which is constant regardless of color, transformation, size, resolution (ideally). This paper usually provides a vector of descriptors for each feature point.

3) Matching: Identify similar features by comparing descriptors in an image. For two images, the result can get a set of pairs (Xi,Yi)->(Xi', Yi') , where (Xi,Yi) is a feature of one image and (Xi' , Yi) is a feature of the other image, because neural networks are very good at feature extraction and other aspects have very significant results, so this paper focuses on the use of CNNs to extract the features of traffic signals.

Since the advent of AlexNet [3], the growth of CNNs has swept the entire computer landscape. Not only does it provide an end-to-end processing approach, but it also significantly outperforms the accuracy of our eyes. To be able to apply our model to vehicles, this work first compared the overhead of several classical CNNs and decided to improve on the Inception module.

Since AlexNet won the ILSVRC 2012 ImageNet image classification competition [3], the convolutional neural network (CNN) boom has swept the entire computer vision field. It not only provides an end-to-end processing approach, but also significantly refreshes the accuracy of each image competition task, and even surpasses the accuracy of the human eye. To be able to apply our model in vehi-

DOI: 10.54254/2755-2721/2/20220531

cles, this work first compared the overhead of several classical CNNs and decided to improve on GoogleNet.

In the VGG network [4] its network structure is similar to ALEXNET, which is to continuously stack convolutional and pooling layers (from bottom to top), VGG proposes a new idea: namely, chunking. The convolutional layers are divided into individual Vgg_blocks, and then a Vggstack function is defined for stacking. In the subsequent GoogleNet and its various versions, Google researchers proposed a network structure that used a very efficient inception module to obtain a deeper network structure than VGG, but with fewer parameters than VGG, because it removed the fully connected layer at the back and also used a four-way parallel structure, so the parameters were greatly reduced, while having very high computational efficiency. The four parallel lines of an inception module are shown on the right The features obtained by the four parallel lines are stitched together in the dimension of channel [5]. In Table 1, this paper summarizes the parameters of the three models:

Table 1. Comparison of the three convolutional neural network models.

Model	Model Size (MB)	Million Mult-Adds	Million Parameters
AlexNet[1]	200	720	60
VGG[2]	500	15300	138
GoogleNet[3]	50	1550	6.8

Table 1 shows the universities, the number of parameters and the number of Mult-Adds of the three convolutional neural networks under our comparison.

The efficiency and speed of the network has been greatly improved based on the previous CNN. However, as the CNN is gradually expanded (deepened), it is found that when the number of model layers increases to a certain level, the effectiveness of the model will decrease instead of increase. In other words, degradation of the deep model occurs. Because when we stack a model, it is logical to assume that the effect will get better and better. However, in fact, this is the problem. "Doing nothing" happens to be one of the hardest things to do with current neural networks. To solve this problem, Kaiming He et al. proposed ResNet, a model whose original purpose was to make the internal structure of the model at least capable of constant mapping to ensure that the network does not degrade at least by continuing to stack as it is stacked [6]. In this paper, the Inception module is improved by adding an asymmetric convolution operation and a residual operation in order to be able to extract the most complete features of the image from different directions.

3.2. RF

For integrated learning, we know that it is by building and combining multiple learators for learning tasks.How to produce a "good and different" individual learner is at the core of integrated learning research.In general, integrated learning can be divided into two major categories: Serial integration methods, which generate the underlying models serially (e.g., AdaBoost [7]). The basic motivation for serial integration is to exploit the dependencies between the underlying models. The performance is improved by giving a larger weight to the misclassified samples. Parallel integration methods, which generate the underlying models in parallel (e.g., Random Forest [8]). The basic motivation for parallel integration is to exploit the independence of the underlying models, since the error can be reduced to a greater extent by averaging. “Feature bagging” is done during the learning process in choosing the random subsetof that feature. These features will be selected in many trees will become interrelated if one or more of the features are strong predictors [9].

DOI: 10.54254/2755-2721/2/20220531

This paper use the random forest algorithm. As its name suggests, a random forest is a forest in which many decision trees are integrated and combined to predict the final outcome, and in a random forest, each tree model is bagged and sampled for training. Also, features are randomly selected, and finally for the trained trees are also randomly selected. The result of this treatment is that the bias of the random forest increases very little, and the variance is reduced due to the averaging of the weakly correlated tree models, resulting in a model with small variance and small bias. Since the random forest can perform well on the dataset, the introduction of two randomnesses makes the random forest not easy to fall into overfitting, and it can detect the mutual influence between FEATURES during the training process. In selecting hyperparameters for random forest algorism, random search outperforms over grid search to minimize the time needed to find models that are good, reaching optimization. The Gaussian process analysis from hyperparameters to validation set performance shows that only some hyperparameters really matter for most datasets, but not too few for practical applications, because different hyperparameters have their own uses. [10].

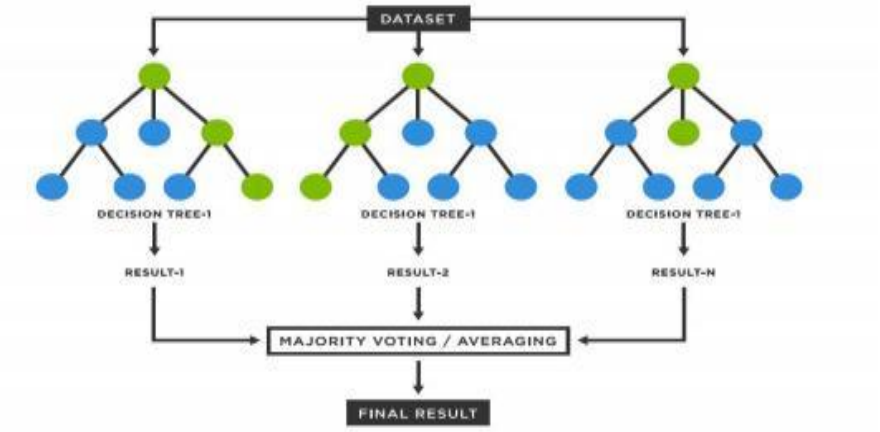

Figure 1. A common RF structure[11]

Figure 1 is the general random forest structure used in most papers

3.3. Overall structure and data processing

3.3.1. Preparation. At the beginning of the code, first let the device can run old version Tensorflow code. Then choose and monitor the GPU the training process used. Them set the size and label of images. Set the input size 48, the root direction of the data, the characteristic number of target label for example ”00000”. Then set the data set, and read each of the image then see which folder this image belongs. Since the images are all RGB type, changing them to gray-scale. After setting the data set, shuffle the training data. Then set the feature list X and label list Y from the training data. Store the data as float32 because float64 is too large to store so many data. Then split X and Y into train and test set. This paper use three popular standards to evaluate the structure.

3.3.2. Enhanced CNN Model. This paper adds a feature extraction structure (PT-Structure) for traffic signatures to improve the original one. It initially has a 5*5 convolution as the first layer of feature input, and extracts the horizontal and vertical features through the convolution kernel of 1 multiplied by 5 and 5 multiplied by 1, and merge them together. Apart from this asymmetric convolution, it also has 5 multiplied by 5 convolution to extract the main feature of the image.And it also has some 1 multiplied by 1 which can reduce parameters and increase the depth of the network.

DOI: 10.54254/2755-2721/2/20220531

Explain the PT-structure with inputting 43 feature maps. Split the 43 feature maps into four branches a, b, c and d respectively (5 multiplied by 5 is called a branch, 1 multiplied by 5 is called b branch, and 5 multiplied by 1 is called c branch and another 5 multiplied by 5 is called d branch), this paper sets the number of a branch output channels to 10, the number of b branch output channels to 15, the number of c branch output channels to 15, and the number of d branch output channels to 10. The branches are independent convolution, The feature can be extracted more diversely.

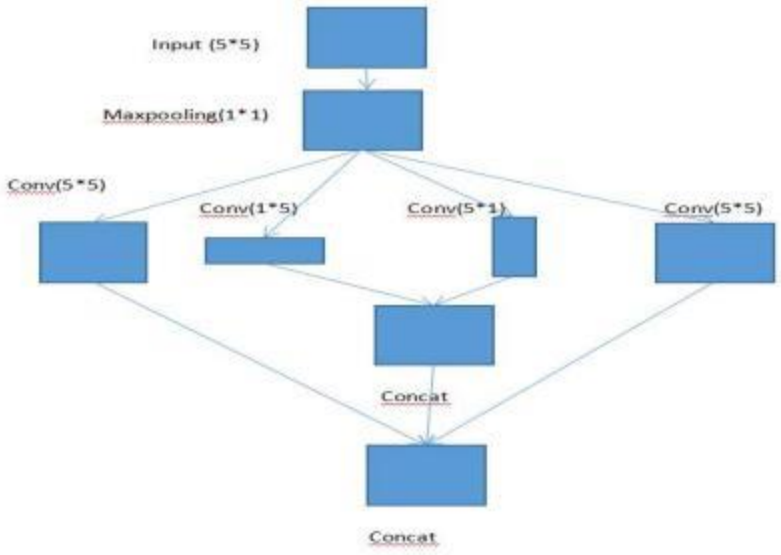

Figure 2. PT structure

Figure 2 is our own neural network structure is built to satisfy both feature extraction and parameter optimization by multi-branching and dimensionality reduction methods

This paper also adds residual structure between layers to extract the largest feature to avoid the vanishing gradient problem. There is batch normalization between each convolution to speed upcompared with the only linear convolution. This paper make a reference (X) input to each layer easier to optimize and can greatly deepen the number of network layers. For our residual structure, the Wi represents the weight of the layer i， where σ represents the non nonlinear function ReLU.

$$F = w_i + 1\ \sigma(w_i x) \tag{1}$$

Then the output y is obtained by a shortcut, and a 2nd ReLU

$$y = F(x, \{w_i\}) + x \tag{2}$$

For our model, x is enough to require another dimension transformation.

DOI: 10.54254/2755-2721/2/20220531

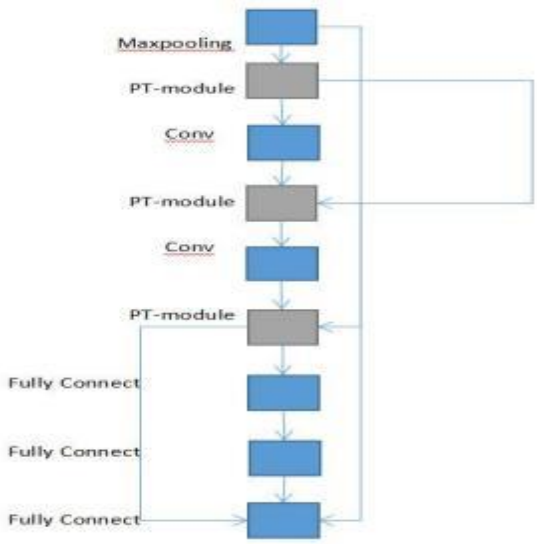

Figure 3. Whole structure

Figure 3 is the main neural network architecture used in this paper, including the superposition of several neural network layers and the residual structure.

3.3.3. Random Forest. After the CNN this paper use random forest to replace full-connected layers. First fit the CNN model, then pick the fully-connected layer just above the softmax. Then fit the extracted features with the labels to random forest. The random search has been used for cross validation. First combine the training and testing data for cross validation, then set estimators randomly for the number of sub-tree, the number of trees in random forest, the max depth for every tree, and the search grid. Then it shows the result.

4. Benchmarks, evaluation and databases

This paper used the German GTSRB dataset, and in addition to the original data, we also added additional noise to the dataset, such as rotating the images, changing their contrast, and adding noise points. In addition, the work also added some images of US traffic signals and Chinese traffic signals to make the set more diverse in features.

This paper also compare the performance of our algorithm with linear regression, general neural network, convolutional neural network-only method, and random forest-only method on this dataset, and finally we find that our algorithm has better generalization and can achieve more than 97% accuracy in a shorter time.

DOI: 10.54254/2755-2721/2/20220531

Table.2 Comparison results of various algorithms.

Model name	Accuracy	Training time
Linear Regression	67 3%	30min
NN	91%	1.5h
CNN	94%	2h
CNN(PT)	97.4%	1.5h
CNN(PT)+RF	97.9%	1.5h
CNN(PT)+RF+Residual	98.2%	1h

Table 2 is the result of this article by comparing different methods.

5. Conclusion and future work

In this project, there are still some problems need to be overcome. The first is the problem of database. The pictures in there have low resolution and inconspicuous feature points, so the accuracy will be impacted. Also these pictures have high similarity which will lead to the training set shows great; however, the performance of test set be weak. The second problem is about the running time. This project takes an hour to run. So the parameter will be adjusted and the Bayesian Optimization will be added into this project in the future. These improvements will enhance the project. In this project many of the structures were used such as the 3-layer CNN, the Random Forest, the Paratroopers-Structure, and the Residual. Each of structure shows great accuracy, but the highest is the combination of the Paratrooper-structure, the Residual and the Random Forest which have the accuracy of 98.3%. Finally in this project, the combination of CNN and random forest has stronger generalization ability in the realization of image recognition. Then, the structure defined by us can effectively extract both horizontal and vertical features, thus achieving high accuracy. Also, by adding multiple fully connected layers and dropout operations at the end, and that can reduce the number of output parameters while ensuring that the output features will not decay.

Reference

[1] Wenhui Li1, Daihui Li1 and Shangyou Zeng1, IOP Conference Series: Materials Science and Engineering, Volume 688, Issue 4, Traffic Sign Recognition with a small convolutional neural network, 2019

[2] Ayoub ELLAHYANI, Mohamed EL ANSARI, Ilyas EL JAAFARI, Said CHARFI, Ibn Zohr University, (IJACSA), Traffic Sign Detection and Recognition using Features Combination and Random Forests, 2016

[3] Alex Krizhevsky, University of Toronto, Ilya Sutskever, University of Toronto ,Geoffrey E. Hinton,University of Toronto , ImageNet Classification with Deep Convolutional Neural Networks,2012

[4] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition , 2015

[5] Christian Szegedy, Wei Liu, Yangqing Jia,(University of North Carolina), Chapel Hill,Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan,(University of Michigan), Vincent, Vanhoucke, Andrew Rabinovich. Going deeper with convolutions ,2014.GoogleNet

[6] Kaiming He， Xiangyu Zhang， Shaoqing Ren， Jian Sun, Cornell University, Deep Residual Learning for Image Recognition, 2015

DOI: 10.54254/2755-2721/2/20220531

[7] Yoav Freund, Robert E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting[J]. Journal of Computer and System Sciences, 1997, 55(1):119- 139.

[8] Leo Breiman， Statistics Department, University of California, Berkeley, Random Forests, 2001

[9] Ho, Tin Kam, ADataComplexityAnalysisofComparativeAdvantagesofDecisionForest Constructors", Pattern Analysis and Applications, 2002

[10] James Bergstra, Yoshua Bengio, Departement d'Informatique et de recherche op ´ erationnelle ´ Universite de Montr ´ eal, Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 13 (2012) 281-305

[11] Christopher Bishop, Published by Springer, January 2006, *Pattern Recognition and Machine Learning*