Comparative Experiment and Performance Evaluation of Three Different Models of Traffic Sign Recognition System

Lisheng Wang

Sydney Smart Technology College, Northeastern University at Qinhuangdao,China

Abstract. Traffic sign recognition is one of the key tasks of intelligent transportation system, which is of great significance to improve traffic safety and efficiency. As deep learning technology continues to advance, more and more studies begin using deep learning models to improve the accuracy of traffic sign recognition. Aiming at the traffic sign recognition task, this essay uses the GTSRB dataset to construct three deep learning models, CNN, MobileNet and ResNet. The experimental results show that in traffic sign recognition tasks, ResNet achieves the highest accuracy of 0.9970, leveraging residual learning for superior feature extraction, making it ideal for high-precision requirements. CNN follows with an accuracy of 0.9827, showing a good balance between performance and simplicity, while MobileNet, though less accurate at 0.7540, excels in efficiency and lightweight design, making it suitable for resource-constrained environments. These methods are designed to address current limitations and further advance the development of traffic sign recognition technology.

1 Introduction

As autonomous driving and intelligent transportation systems become more advanced, traffic sign recognition has become one of the key tasks. Precise detection of traffic signs plays a vital role in ensuring driving safety and boosting traffic efficiency [1]. Developing an efficient and accurate traffic sign recognition system is of great significance for autonomous vehicles, intelligent traffic monitoring and driver assistance systems. By improving the accuracy and robustness of traffic sign recognition, road safety and traffic fluency can be significantly improved.

At present, significant progress has been made in the field of traffic sign recognition. Traditional machine learning methods are widely used in early research, but as deep learning technology matures, convolutional neural networks have increasingly become the dominant method. In recent years, lightweight networks and deep networks have also been widely used in traffic sign recognition tasks [2]. However, these models still face a dilemma between accuracy and computational efficiency, especially when operating in resource-limited settings.

202219192@stu.neu.edu.cn

In this study, GTSRB was selected as the dataset, and normalization and data augmentation were employed to enhance the model's generalization capabilities. Then three distinct models—CNN, MobileNet, and ResNet—on this dataset are constructed and trained. Each model's performance was evaluated with accuracy being the main focus of measurement, with an analysis of the relationship between accuracy and training duration. A detailed comparison of the three models was conducted to assess their effectiveness in traffic sign recognition. Finally, we summarized the experimental results, highlighting the strengths and limitations of each model, and proposed potential directions for future research.

2 Methodology

This paper will use CNN, ResNet, MobileNet three models for the training and recognition of traffic signal signs.

2.1 CNN

Convolutional Neural Networks (CNNs) are a type of deep learning architecture extensively applied in areas such as computer vision, image recognition, speech recognition, and beyond [3]. By simulating the working mechanism of biological vision system, it can automatically extract features from images, so as to realize applications including image classification, object detection, and segmentation. Its multi-layer structure can learn features from simple to complex to better recognize traffic signs (Figure 1).

The working principle of CNN can be distilled into the following steps [4]:

(1) Input image: An image is fed into the network, which can be a grayscale image or a color image.

(2) Convolution operation: In the convolution layer, the convolution kernel traverses the input image, calculating the dot product with local regions to create a new feature map.

(3) Activation function: The activation function is applied to the convolutional feature map to introduce nonlinear factors and enhance the expression ability of the network.

(4) Pooling operation: The pooling layer compresses the spatial dimensions of the feature map, retaining essential feature information while lowering computational requirements.

(5) Repeat convolution, activation, and pooling: Depending on the design of the network, the above steps are repeated to extract more advanced features layer by layer.

(6) Fully connected layer: The extracted features are flattened and input to the fully connected layer. By learning weights and biases, the final classification result or regression value is output.

(7) Loss function and Backpropagation: The loss function measures the gap between the predicted and true values, and subsequently, the backpropagation algorithm fine-tunes the network's weights and biases to improve its performance.

The convolution operation is the most critical part of CNN and its formula is as follows:

$$\text{Output}(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \text{Input}(i+m, j+n) \times \text{Kernel}(m,n) + \text{Bias} \quad (1)$$

The element at location (*i,j*) in the output feature map is given by Output(i,j). The pixel value of the input image at the coordinates (i+m, j+n) is indicated by Input(i+m, j+n); Kernel(m, n) is the convolution kernel at position (m, n); M denotes the height and N denotes the width of the convolution kernel; Bias is usually a scalar..

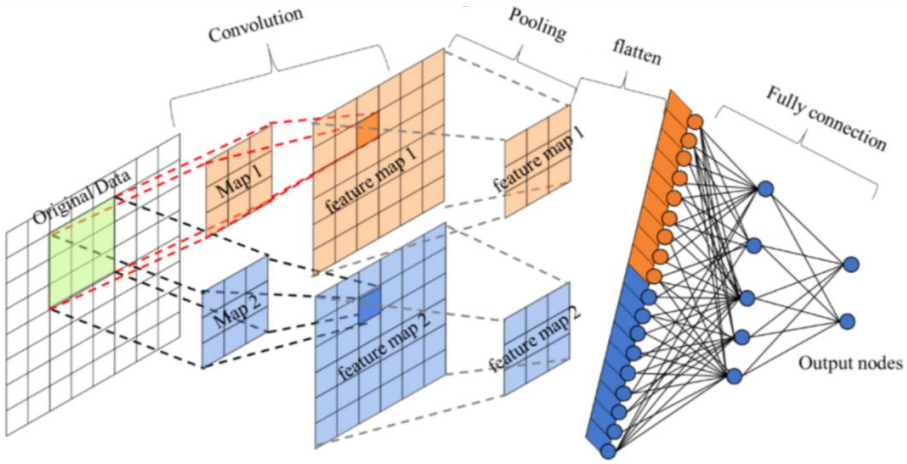

Figure 1 CNN Model diagram [5]

2.2 ResNet

ResNet, also known as Residual Network, was developed by He Kaiming et al. in 2015 as a deep convolutional neural network architecture. By introducing the concept of Residual Learning, It mitigates the issues of gradient vanishing and gradient explosion that arise from excessive depth in traditional deep neural networks [6]. The key idea of ResNet is to introduce Skip connections that allow inputs to skip one or more layers directly, thus optimizing the training process of the network. ResNet simplifies the optimization problem and enables the network to converge to the optimal solution faster. This means that ResNet exhibits faster learning speed and higher efficiency during training

The key principle behind ResNet is described below:

$$H(x)=F(x)+x \qquad (2)$$

(1) H(x) is the target map of the network, which is the output of the ResNet.

(2) F(x) is the residual function that represents the amount of variation in the input at a certain layer of the network.

(3) x is the input and is added directly to the output via skip connections.

In the Residual Block, the input x undergoes two convolution operations to obtain F(x), and then F(x) is summed with the input x to produce the final output. If the dimensions of the input and output are not consistent, they are usually dimensioned by a 1×1 convolution (Figure 2).

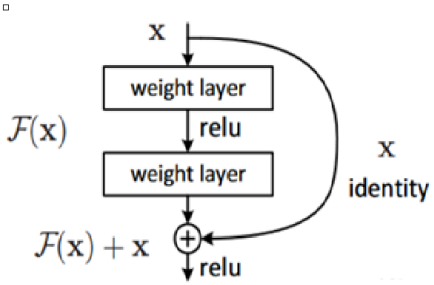

Figure 2 ResNet model diagram [6]

2.3 MobileNet

MobileNet is a lightweight deep convolutional neural network proposed by Google, which aims to optimize the use of computing resources so that it can run efficiently in mobile devices and embedded systems. The core innovation of the network is Depthwise Separable Convolution [7]. The decomposition of standard convolution into depthwise and pointwise convolutions leads to a substantial reduction in computation and model parameters. MobileNet is known for its lightweight and efficient nature, which makes it well suited to run on mobile devices and embedded systems. Traffic sign recognition usually requires real-time processing in resource-constrained environments, and MobileNet's small size and low computational complexity make it an ideal choice (Figure 3)

Taking MobileNet V3 as an example, the hierarchical results are as follows:

(1) Initial convolutional layer: A standard 3×3 convolutional layer.

(2) Improved inverted residual module: multiple modules are stacked, each module includes:

1×1 pointwise convolution: Applied to broaden the channel count.

Depthwise Convolution: Convolution is performed on each input channel independently.

Squeeze-and-Excitation (SE) module: For channel attention optimization to enhance feature representation.

Linear Bottlenecks: Reduce information loss.

Residual join: If the dimensions of the input and output are the same, the input is directly added to the output.

(3) Fully connected layers: Outputs for classification or other tasks.

(4) Softmax layer: Outputs the final classification probabilities.

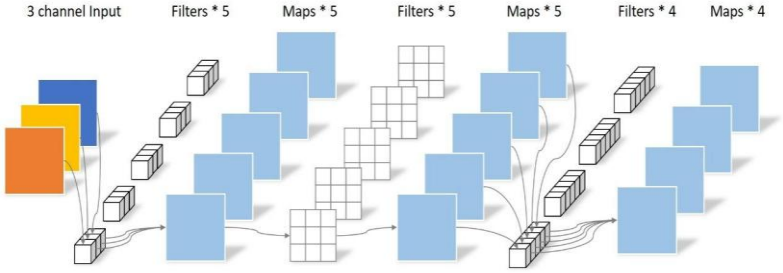

Figure 3 MobileNet model diagram [7]

CNN is a classical convolutional neural network architecture, which has a simple structure and is suitable for a number of vision tasks. However, as the network is deepened, it tends to suffer from gradient-related challenges, which limits the performance improvement. ResNet addresses the training challenges of deep networks through residual learning and skip connections, enabling the training of extremely deep networks, which significantly enhances the model's performance and generalization, though at the cost of high computational and memory demands. MobileNet is based on depthwise separable convolutions, which drastically reduces computation and model parameters, making it suitable for deployment in mobile devices and embedded systems. While the accuracy may be slightly lower than ResNet, its efficiency and lightweight design make it ideal for resource-constrained scenarios

3 Experiment

3.1 Introduction to the Experimental Dataset

The German Traffic Sign Recognition Benchmark, known as GTSRB, is a commonly utilized dataset in the field of traffic sign recognition research. It provides an important basic resource for the research of automatic driving, intelligent transportation system and computer vision. The dataset comprises 43 classes of traffic signs, with over 50,000 images in total, specifically partitioned into 39,209 images for training and 12,630 images for testing [8]. These images, taken under different lighting, weather conditions, and background complexities, exhibit significant diversity and effectively mimic real-world traffic scenes. Each image is accompanied by detailed annotation information, including the category and location of the traffic sign, which makes it well suited for supervised learning tasks to help researchers train and evaluate traffic sign recognition algorithms.

One of the main characteristics of the GTSRB dataset is its high diversity. The images cover common traffic signs in Germany with rich backgrounds and diverse lighting conditions, which provide rich samples for model training and help enhance the model's generalization capacity[9]. In addition, the clear labeling information enables researchers to easily perform data preprocessing and model training, so as to carry out research work more efficiently. The accuracy of the model was used as the experimental index, the relationship between the accuracy and training epoch is plotted as line graphs[10].

3.2 Analysis of Experimental Results

In CNN model, the accuracy gradually increases from 0.1923 in the first cycle to 0.9827 in the 20th cycle, which indicates that the model's performance on the training set is consistently improving. Especially in the first few cycles, the accuracy is improved significantly. The training loss decreases from 2.9764 in the first cycle to 0.0559 in the 20th cycle, which demonstrates that the model is improving its fit to the training data over time, and the error continues to decrease. The decreasing trend of the loss function matches the increasing trend of the accuracy, which is in line with the normal training process (Figure 4).

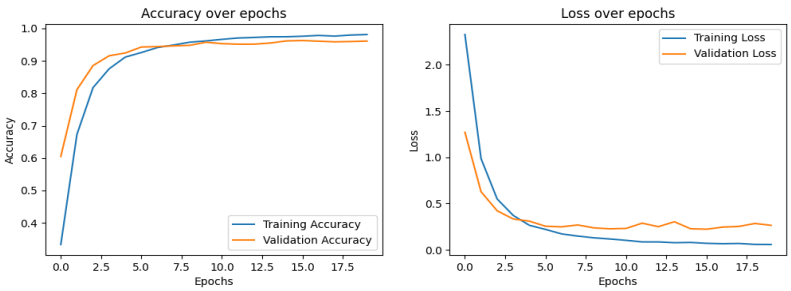

Figure 4 Line graph of CNN results (Picture credit : Original)

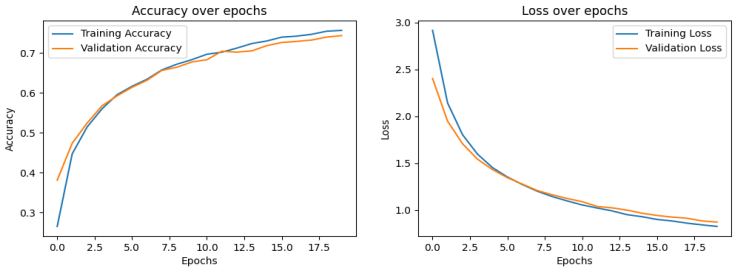

Figure 5 Line graph of MobileNet results (Picture credit : Original)

The accuracy of MobileNet shows a steady climb, beginning at 17.17% during the initial training epoch and reaching 75.40% by the 20th epoch. Despite this enhancement in accuracy, the rate of improvement is notably gradual, with particularly minimal gains observed in the later epochs (Figure 5).

In ResNet model, due to the large number of data sets, it is difficult for data sets to pass through the network at one time, so this paper adopts batch processing of data sets. After many tests, when batch_size is set to 16, that is, 16 images are in a group, the training rounds are 100, and the initial learning rate is 0.01, the training effect is the best (Figure 6).

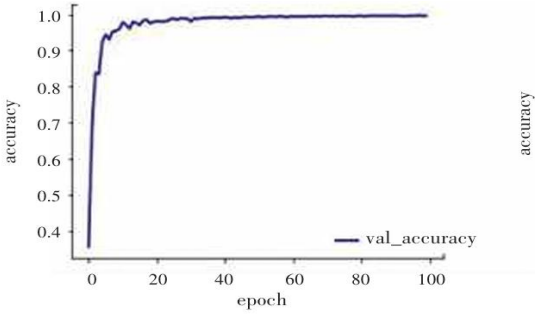

Figure 6 Line graph of MobileNet results (Picture credit : Original)

Table.1 Accuracy of each model

Method	Accuracy
CNN	0.9827
MobileNet	0.7540
ResNet	0.9970

In traffic sign recognition tasks(Table 1), ResNet achieves the highest accuracy of 0.9970 due to its residual learning and skip connections, which enable deeper network training and richer feature extraction, making it ideal for high-precision requirements. CNN follows with an accuracy of 0.9827, leveraging its relatively simple structure for efficient training and high performance in resource-rich environments. However, its accuracy is slightly lower than ResNet's because it lacks the advanced architectural features needed to handle deeper networks effectively. MobileNet, while less accurate (0.7540), excels in efficiency and lightweight design, making it suitable for resource-constrained devices like mobile phones. Its reduced accuracy stems from its use of depthwise separable convolutions, which optimize computational efficiency but sacrifice feature extraction capability in complex tasks. Overall, ResNet outperforms the other models by balancing depth and efficiency, while CNN and MobileNet offer trade-offs between accuracy and computational demands.

4 Conclusion

Aiming at the traffic sign identification task, this paper uses the GTSRB dataset, and constructs three deep learning models of CNN, MobileNet and ResNet for experiments. Through the Python and TensorFlow framework, the training and evaluation of the three models are realized, and the accuracy is used as the main evaluation index, and the relationship between accuracy and training epoch is plotted.

The experimental results indicate that in traffic sign recognition tasks, the CNN model performs effectively, and the ResNet model achieves the highest accuracy of 0.9970. However, the MobileNet model, while optimized for efficiency, exhibits relatively low accuracy, partly due to reduced feature extraction capability in complex tasks.

For future work, several directions are proposed. First, model optimization could focus on further enhancing MobileNet, potentially by integrating a more efficient depthwise separable convolutional structure or incorporating beneficial elements from other lightweight models to boost its performance in complex scenarios. Second, employing more advanced approaches to expand data volume and diversity, such as random cropping and color transformation, could significantly improve the model's generalization ability. Lastly, exploring multi-model fusion methods that combine the strengths of CNN, MobileNet, and ResNet could lead to higher accuracy and enhanced robustness in traffic sign recognition. These strategies aim to address the current limitations and further advance the field of traffic sign recognition technology.

References

1. Ren, J., Huang, R. N., "Current Trends in Deep Learning for Autonomous Vehicles: A Comprehensive Review," J. Eng. Res. Sci., vol. 1, no. 10, pp. 56–68.
2. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C., "MobileNetV2: Exploring Inverted Residuals and Linear Bottlenecks for Efficient CNN Design," 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 4510–4520.
3. Shorna, S.A., Ayon, W.I.Z., and Apu, K.I.Z., "Leveraging CNN Architectures for Enhanced Traffic Sign Classification: Insights from the GTSRB Dataset," 2024.
4. Liu, Z. Li, F., Yang, W., Peng, S., and Zhou, J., "Survey of Convolutional Neural Networks: Key Analysis, Applications, and Future Directions," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019.
5. He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Networks for Image Recognition," CVPR, 2016.
6. Zhu, Y., Zuo, Y., Zhou, T., and Fan, G., "Design of a Multi-Mode Visual Recognition Hardware Accelerator for AR/MR Glasses," 2018 IEEE Int. Symp. Circuits Syst., pp. 1–5.
7. Yurtsever, E. and Lambert, J., "Emerging Technologies and Best Practices in Autonomous Driving: A Comprehensive Survey," IEEE Access, vol. 8, pp. 58443–58469.
8. Sun, Q. and Luo, X., "Palm Vein Recognition Using Transfer Learning and MobileNetV2 Model," 2022 4th Int. Conf. Front. Technol. Inf. Comput., pp. 559–564.
9. Guo, Y., Feng, W., Yin, F., and Liu, C. L., "Signparser: An end-to-end framework for traffic sign understanding," International Journal of Computer Vision, 132(3):805–821, 2024.

10. Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C., "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," Neural Networks, 2012.