**Attention-aware CNN model for Traffic Signs Classification**

**Xiangsheng Gu1a\***† **Qiuyi Lang2b\***† **Fu Lin3c\***† **Pengfei Wang4d\***†

1Donald Bren School of Information & Computer Science, University of California, Irvine, USA

2Computer Science, Queen’s University Belfast, Belfast, UK

3Lassonde School of Engineering, York University, Canada

4Leicester International Institute, Dalian University of Technology Panjin, China

\*Corresponding author. Email: axiangshg@uci.edu; b[qlang01@qub.ac.uk;](about:blank) c[lf52138@my.yorku.ca;](about:blank) dpw185@student.le.ac.uk

†These authors contributed equally.

**Abstract**. Autonomous vehicle driving systems have become one of the most important topics recently, some people assert that they can identify traffic signs automatically, which is a revolutionary improvement for transportation; however, the accuracy of autonomous vehicle driving systems still remains controversial. Therefore, this research analyzes the German Traffic Sign Benchmark dataset, in three different ways: AlexNet, VGG-16, and ResNet-50 of autonomous vehicle driving systems, to compare the best approach for identifying traffic signs. One can conclude that AlexNet, VGG-16 and ResNet-50 performed well, as they got an accuracy score of 95%, 95%, 89% respectively. To improve the classification accuracy to achieve nearly 100% for total security, Bottleneck Attention Module (BAM) is examined as a way of improving the classification accuracy of models and it is confirmed that BAM is able to boost the accuracy score of select models.

**1. Introduction**

Research on how to reduce traffic fatalities has been around for years (Evans, Leonard, 2004)[1]. Since 1980, various traffic sign recognition systems have been developed. For instance, one focused on real time traffic sign detection [2] and one engaged in dealing with object recognition in outdoor environments [3]. With the accelerating process of modernization, the ownership of cars has been an increasing trend since the last half century, it has played an important role in technological advancements. Numerous works have been proposed on data recognition. In reference [4], although the existing traffic sign recognition method based on convolutional neural networks can improve the accuracy rate by increasing the number of network layers, there are problems such as reduced efficiency, increased parameter quantity, and long training time. Huo et al. improved the SqueezeNet model [4]. They used the ELU function as the activation function to improve the learning efficiency, introduced a deep residual network to avoid the gradient disappearance when the network was too deep, and used the GRU neural network to memorize the past important features. Memory to ensure the stability of the model. In the end, they improved the recognition accuracy and recall rate of traffic signs, shortened the training time, and improved the convergence speed and stability. When detecting traffic signs, due to their small and dense characteristics, they [5] are easily affected by the complex natural environment, resulting in poor detection performance, as well as the need to ensure driving safety. It is necessary to conduct real-time detection of small-scale signs in the distance in advance. These conditions require the detection network not to lose the details of small targets after deep feature extraction, but also to have a strong ability to represent weak target features, while taking account of the real-time performance. Li et al. proposed Attention based Multi-scale Small Target Traffic Sign Detection (AMST-TSD), which uses CSPDarknet53 to extract features, and optimizes the network structure according to the scale characteristics of traffic signs, and adopts deconvolution adaptive cascade. In this way, the detailed information of the shallow network is preserved, an inverted pyramid structure based on the spatial attention mechanism is designed. The detection performance of saliency regions of low-resolution feature maps are enhanced by using the high-resolution feature map. Finally, the performance of the algorithm is significantly improved while ensuring real-time performance, and the impact of harsh natural environments on traffic sign detection is effectively solved. In general, the model has performed well when it comes to classifying traffic signs in different scenarios. In harsh environments [6], the application of automatic classification technology of traffic signs is still insufficient, especially in winter when snowy weather, lighting, occlusion, and other unpredictable factors make related computer vision more difficult.

Zhou et al. proposed a parallel fusion attention network (PFANet) based on high-resolution traffic sign classification. The network extracts features more effectively by combining two attention modules PFAN-A and PFAN-B, and the module they designed Conducted ablation research. Compared to HRNet-w18, EfficientNet and MicroNet, their proposed attention network achieves the best results on the corresponding ice and snow road sign datasets, as well as on GTSRB.

Nevertheless, the increasing frequency of crash risk cannot be ignored as well. Based on the China Statistical Yearbook (2019), the figure shows that the number of traffic accidents happening in China has had an exponential growth trend in the past five years. Recent theoretical developments have revealed Autonomous vehicle driving systems might become very popular in the future. The autonomous vehicle will be accessing information on the road and react accordingly, however, there are various unpredictable fluctuations that remain controversial in the real world, such as the accuracy of identification and the speed of identification. Different aspects of the traffic signs are characterized and varied approaches to identifying the signs are applied to improve the accuracy of identifying said traffic signs. The aim of this work is to research the actual performance and utilization values of traffic signs. This paper is organized as follows: Section 2 presents the literature review, section 3 contains the data collection and data analysis, section 4 shows the results and discussion, and lastly, section 5 makes conclusions and future extension.

**2. Method**

*2.1 AlexNet*

According to a study done by (Md Zahangir Alom et al., 2018) [7] AlexNet architecture makes use of spatial correlation to improve the accuracy of the ImageNet challenge. Figure 1 shows the AlexNet architecture (Hassan, M.).

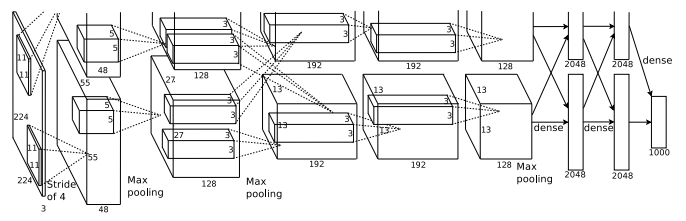


Figure 1. Framework of AlexNet.

It is made of five convolutional layers and three fully connected layers. On the one hand, it used Rectified Linear Unit (ReLU) as activation which deals with the vanish gradient problem due to ReLu's unbounded feature. On the other hand, in order to prevent situations like learning variables becoming high, it introduced Local Response Normalization (LRN) to keep learned variable weight ground. In addition, the AlexNet architecture did data augmentation and dropout to prevent overfitting.

*2.2 VGG16*

VGG16 architecture(Qassim, H., Verma, A., & Feinzimer, D., 2018) [9] was built to meet the requirement of reducing the parameters in convolutional layers. Figure 2 shows the VGG16 architecture (Hassan, 2018).

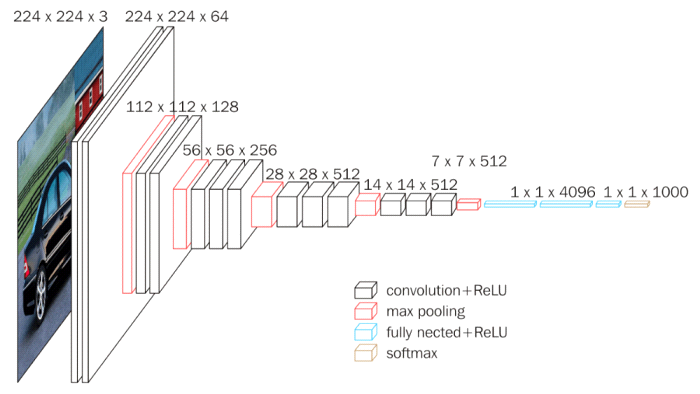


Figure 2. Framework of VGG16.

Compared with AlexNet which consists convolutional layers, all of VGG16 convolutional layers are and get the same effect as AlexNet's different Convolutional layers. As Aqeel Answar discussed [11], VGG16 can have the same output as AlexNet, given an input layer , convolutional layers and stride one will lead to output. Similarly, if there are 2 times conv layers, the identical result would be the output. Therefore, VGG16 decreased the parameters in convolutional layers by replacing all convolutional layers of kernel size to and that resulted in low overfitting. (Yu, W et al., 2016, p5) [12]

*2.3 ResNet*

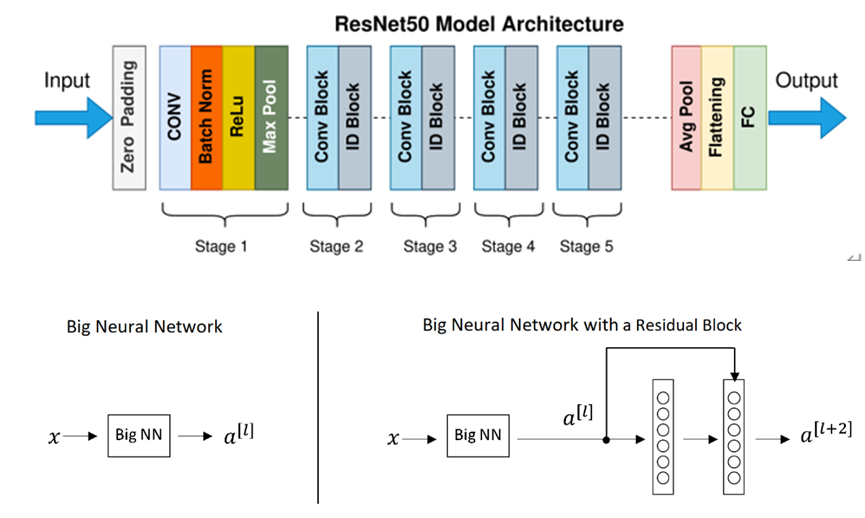


Figure 3. Framework of ResNet.

Wu, Z., Shen, C., & Van Den Hengel, A. (2019) [13] has found that in the deep learning area, it has been proved that more layers can lead to worse results. In addition, the deeper the Convolutional Neural Network (CNN), the less impact it can have on the initial layers. To address these problems, ResNet introduced two types of "shortcuts": the identity shortcut and the projection shortcut. Figure 3 shows the whole ResNet architecture and how these "shortcuts'' work (wikimediacommons, 2021).

Instead of learning from , ResNet learns from . When input and output have the same dimension, which is the identity shortcut and when the size of input and output are different, changing the input size to output size is the projection shortcut.

**3. Results and discussion**

*3.1 Dataset description*

Datasets can have important impacts on test results, thus how to pick the most appropriate dataset is worth considering. First of all, greater diversity and choice of traffic sign is necessary, such as speed limit, vehicle limit, emergency marks and so on. Then, many factors can affect the display of traffic signs in the real world as well, for example identifying traffic signs with different distances, different perspectives, rotated to small angles and some extreme weather. These factors change displayed color representations according to lighting, angles, and brightness. The German Traffic Sign Benchmark [14] is chosen as the dataset for the project, since it has more than 40 classes and over 50,000 images in total. In addition, those pictures are large enough to train and all of them are from real cities in Germany with random fog marks and random snow.



Figure 4. Data sample of the utilized dataset.

*3.2 Comparison results*

Table 1. Results

|  |  |  |
| --- | --- | --- |
| Methods | Training loss | Test accuracy |
| AlexNet | 0.016 | 95% |
| VGG-16 | 0.029 | 95% |
| ResNet-50 | 0.961 | 89% |
| ResNet-50 + BAM | 0.021 | 96% |
| AlexNet + BAM | 0.014 | 97% |
| VGG-16 + BAM | 0.017 | 97% |

This table compared the test accuracy and training loss tested on different methods. In the beginning, both AlexNet and VGG-16 got 95% accuracy and low train loss: 0.016% and 0.026%. ResNet-50 got a relatively low performance by comparison, which is 89% in accuracy and 0.961% in training loss. When we applied BAM to these methods, all accuracy of models improved. AlexNet and VGG-16 resulted in a 2% increase than the original ones, and ResNet also got an increase of 7% than previous results. Overall, BAM plays a positive role in improving classification precision on all models and significantly increased accuracy on ResNet the most.

**4. Conclusion**

Three well-known pre-trained networks are applied to classify traffic signs. First, image features are extracted by using a pre-trained network and then extracted features are put as input to fully connected pre-built layers. The precision rate is improved by introducing data augmentation and Fine-tuning to avoid over-fitting. In practice, results are in the range 89%~97%. As Table 1 presented, the Model based on ResNet-50 got a performance of 89% and the model using VGG-16 boosted the precision rate up to 95%. Furthermore, the accuracy got 97% after using AlexNet to extract features. In order to improve the performance, The Block Attention Module(BAM) is introduced and applied to the models. BAM is designed to efficiently improve intermediate features by learning what and where to focus on, or suppress through two branches, channel attention and spatial attention, as well as ultimately obtain the final attention mapping by element-wise summation. The result showed the classification accuracy slightly improved in both AlexNet and VGG-16 after applying BAM, as the precision rate only went up from 95% to 97%. Also, one thing worth noting is that it took a shorter time to train AlexNet compared to the other two modified models. However, ResNet-50 achieved an outperformed result, with 97% correct classification rate after use of BAM, and accuracy is refined by up to 7% when applying BAM to improve the middle features.

In practice, although both ResNet-50 and VGG16 models achieved outstanding performance, it is difficult to say that BAM can make improvements on any model's classification accuracy. Also, as long as the accuracy score is not 100%, doubts on whether they can well identify the traffic sign when they worked in real life will still remain. This is because of the surrounding complex environmental influences, like uncontrolled weather conditions and unpredicted traffic environments. Therefore, we believe it is key to collect more data to train models, as the more data is collected, the more reliability and precision these models can achieve.

**References**

1. Evans, L. (2004). *Traffic safety*.
2. Estable, S., Schick, J., Stein, F., Janssen, R., Ott, R., Ritter, W., & Zheng, Y. J. (1994, October). A real-time traffic sign recognition system. In Proceedings of the Intelligent Vehicles' 94 Symposium (pp. 213-218). IEEE.
3. De la Escalera, A., Armingol, J. M., & Mata, M. (2003). Traffic sign recognition and analysis for intelligent vehicles. Image and vision computing, 21(3), 247-258.
4. HUO Ai qing, ZHANG Wen le, LI Hao ping. Traffic road sign recognition based on SqueezeNet model with deep residual network and GRU[J]. Computer Engineering & Science, 2020, 42(11): 2030-2036.
5. LI Ye, GU Chen-feng. Multi-scale Small Target Traffic Sign Detection Based on Attention Mechanism. Journal of Chinese Computer Systems, 2022, 43(2): 381-386.
6. Zhou, K., Zhan, Y., & Fu, D. (2021). Learning region-based attention network for traffic sign recognition. Sensors, 21(3), 686.
7. Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164.
8. Hassan, M. ul. (2018). AlextNet Architecture. AlexNet – ImageNet Classification with Deep Convolutional Neural Networks. Neurohive. Retrieved June 5, 2022, from <https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/>Estable, S., Schick, J., Stein, F., Janssen, R., Ott, R., Ritter, W., & Zheng, Y. J. (1994, October). A real-time traffic sign recognition system. In Proceedings of the Intelligent Vehicles' 94 Symposium (pp. 213-218). IEEE.
9. Qassim, H., Verma, A., & Feinzimer, D. (2018, January). Compressed residual-VGG16 CNN model for big data places image recognition. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 169-175). IEEE.
10. Hassan, M. ul. (2018). VGG16 Architecture. VGG16 – Convolutional Network for Classification and Detection. neurohive. Retrieved June 5, 2022, from <https://neurohive.io/en/popular-networks/vgg16/>.
11. Answer, A., 2019. Difference between AlexNet, VGGNet, ResNet and Inception. [online] Medium. Available at: <https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96> [Accessed 5 June 2022].
12. Yu, W., Yang, K., Bai, Y., Xiao, T., Yao, H., & Rui, Y. (2016, June). Visualizing and comparing AlexNet and VGG using deconvolutional layers. In Proceedings of the 33 rd International Conference on Machine Learning.
13. Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition, 90, 119-133.
14. S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel (2011), The German Traffic Sign Recognition Benchmark: A multi-class classification competition, in International Joint Conference on Neural Networks, pages 1453–1460.
15. Hassan, M. ul. (2018). AlextNet Architecture. AlexNet – ImageNet Classification with Deep Convolutional Neural Networks. Neurohive. Retrieved June 5, 2022, from <https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/>.
16. wikimediacommons. (2021). File:ResNet50.png. Retrieved June 5, 2022, from <https://commons.wikimedia.org/wiki/File:ResNet50.png>.