Deep Learning-Based Traffic Sign Identification Using Multi-Scale Feature Fusion

Nikesh Devkota
Department of Information and Communication Engineering
Changwon National University
Changwon, Korea
20227085@gs.cwnu.ac.kr

*Byung Wook Kim
Department of Information and Communication Engineering
Changwon National University
Changwon, Korea
bwkim@changwon.ac.kr

Abstract— **Roadside traffic signs are used to enforce vehicle safety laws, appropriately convey traffic conditions, and aid in safe driving. Therefore, accurate traffic sign identification is essential for automated driving. In this article, we suggest a simple and reliable one-stage object detection model for detecting and identifying traffic signs by using ResNet50 and a Feature Pyramid Network (FPN). We evaluate our network using the Tsinghua-Tencent 100K (TT100K) benchmark for traffic sign recognition, and the results demonstrate that the detection network can obtain overall Mean Average Precision (mAP) of 46.012%.**

Keywords—Deep Learning, Autonomous Vehicles, Traffic Sign Detection, Traffic Sign Recognition, Traffic Sign Dataset, Artificial Intelligence

I. INTRODUCTION

Artificial Intelligence (AI) has made tremendous progress in recent years toward practical applications in a variety of fields including automotive industries. The introduction of Autonomous Vehicles (AVs) in automotive industries have begun to transition from laboratory development and testing to driving on public roads [1]. Because most traffic accidents are caused by human error, autonomous vehicles can be used to automate, adapt, and improve vehicle technology for safety and better driving. Traffic sign detectors are one of the most essential components of self-driving cars. Traffic signs on the road are used to regulate transportation rules, vehicle safety, accurately describe road conditions, and provide driving assistance [2]. However, traffic sign detection is a complicated issue because the algorithms must deal with complex dynamic environments, high accuracy demands, and real-time constraints. Some of the problems in a traffic sign detection system include color and shape variations caused by scene illumination differences, orientation differences, and dissent properties caused by wear-off [2].

Significant advances in several fields of machine learning, particularly computer vision, have been made because of recent advances and the performance of deep neural networks. During training, these networks learn to recognize features that are crucial to identifying between various kinds of objects at different scales. Hence, they are one of the most powerful tools that can be used to automatically identify traffic signs in real life scenarios. The common process for identifying traffic signs automatically using convolutional neural network (CNN) is to extract some features for each traffic sign and train a model to locate and classify them using the extracted features. This process requires a huge volume of traffic sign data which we don't currently own. Hence, we rely on publicly available Tsinghua-Tencent 100K (TT100K)[3] benchmark traffic sign recognition benchmark dataset to train our custom deep learning model and identify traffic signs automatically.

The primary contribution of our work are as follows:

- We utilize a ResNet50 [4] architecture pretrained on ImageNet dataset to obtain feature maps at various spatial resolution from input images.
- We propose a feature pyramid network (FPN) to combine different features obtained from a pretrained ResNet50 architectue.
- Finally, we use a classification subnet in parallel with a bounding box regression subnet to identify traffic signs in the input images.

II.PROPOSED MODEL

Initially a ResNet50 [4] architecture pretrained on ImageNet dataset is used to extract multi-scale feature maps from the input images containing traffic signs. Then, an FPN is used to combine the low-resolution feature maps that are semantically rich with the semantically poor but high-resolution feature maps. As a result, the FPN is able to extract multi-scale information from the input images. Finally, a classification subnet in parallel with a bounding box regression subnet is used to identify the traffic signs in the input images.

A. ResNet

When deeper networks begin to converge, a degradation problem arises as network depth increases, accuracy becomes saturated and rapidly degrades [4]. Surprisingly, overfitting does not cause such degradation, and adding more layers to a sufficiently deep model increases training error. The decrease in training accuracy indicates that not all systems are as simple to optimize. ResNet combats degradation by introducing a

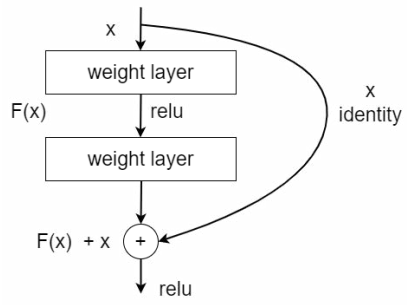

Fig. 1. Residual learning in ResNet

deep residual learning framework. He et al. [4] explicitly allow these layers to fit a residual mapping rather than assuming that each few stacked layers will directly fit a desired underlying mapping. The residual mapping used in ResNet architecture is illustrated in Fig.1.

B. FPN

A key challenge in many computer vision applications, including object recognition and instance segmentation, is the difficulty of detecting objects at various scales in an image. In order to solve this problem, Lin et al. [5] proposed an FPN, which extracts and combines multi-scale features from an image using a bottom-up path and a top-down path. The bottom-up path is a conventional CNN that generates multiple feature maps at various scales. Then, the top-down path creates a feature pyramid by combining multi-scale features obtained from the bottom-up path. In this study, we use ResNet50 as the bottom-up path.

TABLE I. RESNET50 FOR IMAGENET DATASET

Block Name	Output Map Size	Down sampling stride
Conv1	112 x 112	2
Conv2	56 x 56	2
Conv3	28 x 28	2
Conv4	14 x 14	2
Conv5	7 x 7	2

ResNet50 is composed of five convolutional block groups (Conv1 to Conv5), and each block performs down sampling with a stride of 2 [5] as shown in Table I. We utilize the blocks from Conv3 to Conv5 to generate an FPN. Similar to PANET [6], we use a bidirectional augmentation approach i.e., top-down path augmentation followed by bottom-up path augmentation. However, we treat each bidirectional (top-down & bottom-up) path as one feature network layer and repeat the same layer 2 times to enable more high-level feature fusion. Further, we add a skip connection from the original input to output node at the P4 block. The obtained feature pyramid maps at different levels are named {P3, P4, P5, P6, P7} where Pl denotes the feature pyramid maps at l level. If the input image size is 1024 x 1024 pixels, then P3 will have a feature map of size 128 x 128 pixels and P7 will have a feature map of size 8 x 8 pixels. The proposed FPN used in our traffic sign detection model is shown in Fig. 2.

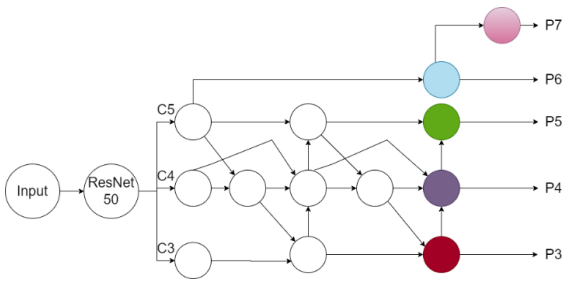

Fig. 2. Proposed FPN for traffic sign detection.

C. Subnet Heads

The subnet head is classified into classification subnet head and box regression subnet head. For each of the A anchors and K object classes, the classification subnet predicts the probability of object presence at each spatial position [7]. The classification subnet applies four 3 x 3 conv layers to an input feature map containing C channels from a given pyramid level, followed by a 3 x 3 conv layer containing K x A filters. Finally, sigmoid activations are used to generate K x A binary predictions for each spatial location [7]. In our experiment, we have set A =9 and C= 256.

The box regression subnet head performs bounding box regression in parallel to the classification subnet head. The box regression subnet is designed similarly to the classification subnet, with the exception that it has four linear outputs per spatial location [7]. For each anchor per spatial location, these four outputs predict the relative offset between the anchor and the GT box.

D. Loss Function

To address the extreme foreground-background class imbalance, the simple and highly effective focal loss strategy [7] is used as the loss on the classification subnet output. We utilize Smooth L1 loss to minimize the bounding box regression loss.

TABLE II. TRAINING PARAMETERS

Parameters	Value
Initial Learning Rate	10^{-5}
Epochs	50
Number of Classes	45
Number of Anchors	9
Input Image Size	800 x 800

III. EXPERIMENTAL RESULTS

The parameters used for training the proposed model have been highlighted in Table II. As in [3], we ignored the data from classes having less than 100 instances resulting in total of 45 target classes. The classes having less than 1000 images are augmented through brightness enhancement, change in image contrast and adding the effect of rain on the original images. During performance evaluation, we computed the Mean Average Precision (mAP) at an IOU threshold of 0.5 using the precision-recall curve. Table III shows the

Fig. 3. Traffic sign detection on sample images

performance comparison of our proposed model with previous work. Our proposed method was able to achieve a mAP score of 0.46.

TABLE III. PERFORMANCE COMPARISON

Model	mAP
Pon et al.	0.31
Cao et al.	0.44
Li et al.	0.33
Proposed Method	**0.46**

Compared with Pon et al. [8] and Cao et al. [9] baseline, our proposed model's mAP score is higher by 0.15 and 0.02 respectively. We also achieved an increase of 0.13 in mAP compared with Li et al. [10]. Traffic sign detection results on sample images are illustrated in Fig.3.

IV.CONCLUSIONS

In this paper, we proposed a feature fusion network for one stage object detection model to detect traffic signs in TT100K benchmark dataset and obtained a mAP of 46.012 %. In future, the performance of our network will be enhanced by incorporating an attention mechanism into the backbone model. In addition, the effectiveness of our algorithm will also be evaluated using other benchmark datasets for traffic signs.

ACKNOWLEDGMENT

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (NRF-2022R1A2B5B01001543).

REFERENCES

[1] S. Grigorescu, B. Trasnea, T. Cocias and G. Macesanu, "A survey of deep learning techniques for autonomous driving. Journal of Field Robotics," vol. 37, no. 3, pp. 362-386, 2020.

[2] Y. Rehman, I. Riaz, X. Fan and H. Shin, "D-patches: effective traffic sign detection with occlusion handling," IET Computer Vision, vol. 11, pp. 368-377, 2017.

[3] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li and S. Hu, "Traffic-sign detection and classification in the wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2110--2118.

[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, 2016.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[6] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980--2988.

[8] A. D. Pon, O. Andrienko, A. Harakeh and S. L. Waslander, "A hierarchical deep architecture and mini-batch selection method for joint traffic sign and light detection," in 2018 15th Conference on Computer and Robot Vision (CRV), IEEE, 2018, pp. 102--109.

[9] J. Cao, J. Zhang and W. Huang, "Traffic Sign Detection and Recognition Using Multi-Scale Fusion and Prime Sample Attention," IEEE Access, vol. 9, pp. 3579-3591, 2020.

[10] L. Cheng-Lin and C.-Y. Su, "Traffic Signs Detection Based on Enhanced YOLOv5 Network Model," in 2022 IEEE International Conference on Consumer Electronics-Taiwan, IEEE, 2022, pp. 449-450.