Fast and Light Object Classification from Aerial Drone Image

***Abstract*—This paper proposes the fast and solid method to classify objects from the aerial drone image with different altitude and perspective angles. In order to classify objects from drone datasets with various characteristics, there are two difficulties. The first difficulty is that the object size is too small. The second difficulty is that the object has many various characteristics. Therefore, we proposed a simple convolution and shortcut blocks with a dilate convolution. Since the drone image is a 4k (3840x2160) in size so thus a very high resolution, neighboring pixels for a pixel of interest are similar to each other, so it is advantageous in terms of speed to use the skipped pixels. However, as the network going deeper, the feature map size becomes smaller and there is a possibility that the surrounding pixel values become much different. Therefore, the normal convolution layer was also arranged and constructed. In addition, since the objects in the drone image have a small size because taken at a high altitude, it loses more features as the network gets deeper. Therefore, the features extracted from the original image were supplemented by periodically calculating each pixel by using the shortcut technique. It records fewer parameters than other networks, thus it executes faster and more accurately. The number of parameters is 1,032,054, which is lower than MobileNet (3,504,872), ShuffleNet (1,366,792), and SqueezeNet (1,248,424). Also, the proposed network outperformed with an accuracy of 98.18%, which is higher than MobileNet (97.28%), ShuffleNet (95.88%), and SqueezeNet (94.32%). when classifying images with 4 GTX3090s, it took in a speed of 1,102 FPS. Therefore, it was proved that fast and accurate network configuration is possible when classification experiments of Aerial Drone Images with various features are conducted using a simple convolution block and shortcut block.**

***Index Terms*—Classification, Atrous Convolution, Drone Dataset, Shortcut**

I. Introduction

Recently, there has been a growing interest in drones worldwide. Among them, the core technologies of drones are autonomous flight and collision avoidance technology, and flight control systems. In order to develop such a system, it is necessary to use the vision sensor mounted on the drone to understand the drone's surrounding environment well. There are no lanes or signs in the sky, and there are many obstacles such as birds, electric poles, and buildings. Therefore, in this paper, we focused on classification among various object recognition techniques using drone view image data. Among them, we focused on reducing the weight of the classification network and improving accuracy. Classification is the most basic part of object recognition technique, and it is the task of classifying object classes. Representatively, GoogleNet [1] proposes the idea of Auxiliary Classifier and ResNet [2] proposes Residual Block to improve the performance by passing the weights and features of the previous feature map. In addition, various convolution methods have been suggested in papers such as [3] and [4].

In this paper, various methods were applied and experiments were conducted. Inspired by the shortcut method used in [5], [6], and [7] and the parameters and methods used to increase the speed in [8], [9], [10], a drone image classification network was constructed. Overall, there are three reasons for this configuration. The first reason is that, due to the nature of the drone dataset, it was filmed at a high altitude, so it was different from the existing data in appearance. In other words, since data photographed from the ground is two-dimensional and data photographed by a drone is three-dimensional, the characteristics of objects become more diversified, and classification accuracy can be improved by approaching them in various ways. The second reason is that even the same object can look completely different depending on the tilting angle, and its characteristics vary a lot. For comparison, imagine a car shot at a 90-degree angle versus a car shot at a 45-degree angle. The third reason is that there are many small objects in size. The drone image used in this paper has a high resolution of 4k, but the size of the object in it is not large. Therefore, the feature that disappears after passing through the convolution layer can damage the accuracy improvement.

Therefore, in this paper, an experiment was conducted using both atrous convolution and Normal Convolution to find various features of each object. In the process, since the object size is small and many features are lost, the Shortcut technique was used to supplement it. Through several experiments, parameter values optimized for the dataset were obtained, and average accuracy of 98.18% was obtained. In summary, the main contributions of this work include:

- We introduce aerial image dataset and explain why this data has various characteristics.
- We propose Convolution Block(CB) including atrous convolution layer and Shortcut Block(SB) to convey feature of original input image. These consists of small number of channels to reduce parameters.
- Our method achieve higher accuracy than state-of-the-art models with lower number of parameters.

II. Related Work

A. Atrous Convolution

There are several types of convolution methods, but in this paper, atrous Convolution is used. This method, also called dilate convolution, has been proposed and used in papers such as [3] and [4]. The main feature is that the Receptive Field can

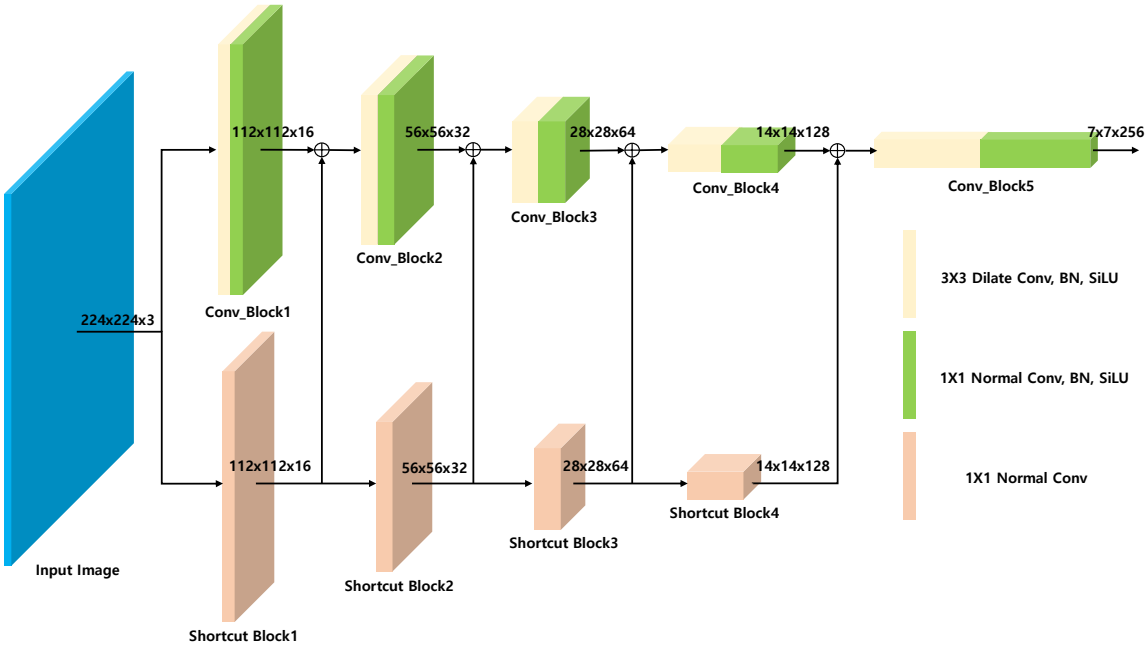

Fig. 1. Proposed Network with Convolution Block(CB) and Shortcut Block(SB).

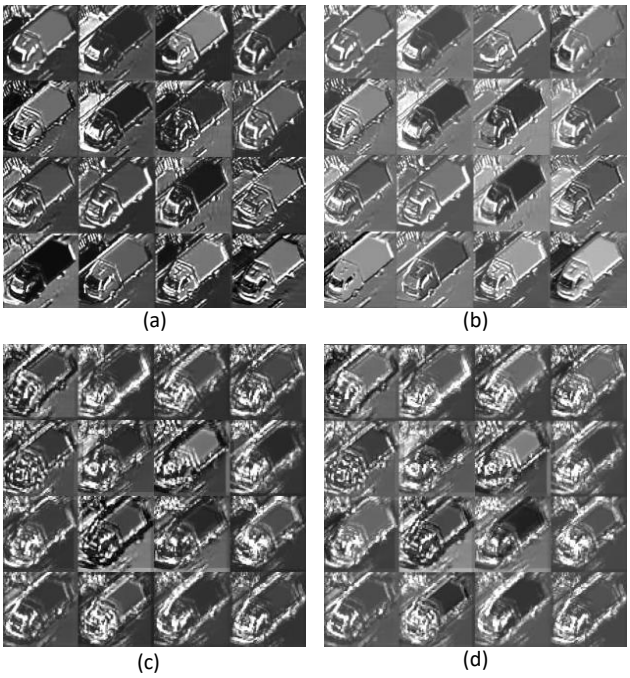

Fig. 2. (a) Output image of first convolution block. (b) Output image of first convolution block with shortcut Block. (c) Output image of second convolution block. (d) Output image of second convolution block with shortcut block.

be increased with a small computational cost. It is mainly used for segmentation work, but in the case of a high-resolution image, the pixel value of one part of the object is similar to or equal to the surrounding pixels, so using atrous convolution, training can be conducted under more favorable conditions.

B. *Shortcut*

Shortcut Connection, also called Skip Connection, has been proposed to solve the problem of gradient vanishing and exploding that appears as the artificial neural network deepens. A representative paper is ResNet [2], which directly transfers the gradient through a simple operation and improves the performance. In a similar way, a shortcut technique was used in [5], a block with reduced number of channels was concatenated, and a high-density network was constructed, and the performance was improved by modifying the inside of the residual block in [11].

III. Proposed Work

The network in this paper is easy to understand as shown in Fig.1. The configured network is structured very sequentially. Simply, CB and SB are connected in sequentially.

A. *Convolution Block(CB)*

The CB consists of one dilate convolution layer and one normal convolution layer. The dilate convolution layer looks like an 11x11 size, but actually it works like an 3x3 size. Because there is a space of 5 pixels between each pixel of a 3x3 kernel. Configure like this, the receptive field is widened and many areas can be viewed at once. After convolution operation, the feature map is stably generated using batch normalization and SiLU activation function. We adopted SiLU because it gives better results than LeakyReLU when the output signal is negative. So far, this is the first part of the CB. The second part of the CB is made by normal convolution layer (1x1 size). And then, there are same situation of BN layer and activation function. After that, the size is reduced to 1/4 through the maxpooling layer. As the calculations proceeds,

Fig. 3. Example Image of Drone Image Dataset.

the characteristics of small objects in the drone image may be lost, resulting in poor results. To solve these problems, we propose a shortcut block.

B. Shortcut Block(SB)

The fundamental purpose of original shortcut is to transmit gradient value, it was created to compensate for the missing features in this experiment. This part conveys the characteristics of the original input image. so, the SB consists of only one normal convolution layer (1x1 size). And this part has no any BN layer and activation function. When used in this way, the number of parameters is reduced, so it is not only faster, but also the characteristics of the original image can be conveyed as closely as possible.

The difference between the feature map with and without the SB can be shown in Fig. 2(a) is the output image after the input image has passed through the first CB. Fig. 2(b) is the output of the CB and the features of the image using the SB give the effect that the features of the surrounding background disappear a little and the features of the object are added to make the object more distinct. The two images below are before Fig. 2(c) and after Fig. 2(d) applying the SB to the second CB. As you can see, the same effect is applied. Using this method, the features of the feature map were further emphasized and the network was configured to increase the accuracy of the experiment.

IV. Experiment

A. Drone Image Dataset

It was used Drone Image for Autonomous Navigation, provided by AIHub. It contains three regions such as downtown, tourism places, and mountains where the image taken in the different altitude and shooting angles while the drone flies below 100m and perspect views up to orthographic views as shown in Fig.3. Thus, variety of objects with different sizes of 6 categories such as cars, trucks, bus and house, tree, etc. Fig.3 shows the examples from downtown and tourism places image extracted from a drone image dataset. Drones can secure a hemispherical vision. If you take an example with a car, you can see the top, including the front and side. Also, if you shoot at a 90-degree angle, you can only see the top of the vehicle, so it becomes a completely different image from the existing data. And in the case of a street lamp, when taken from a distance, it rises perpendicular to the horizon, but as the camera approaches closer, it seems to gradually tilt toward the horizon. If the data is generated in that state, there is a high possibility that it will be difficult to identify the object as the proportion of the street lamp in the image is not large. For this reason, a mixture of Dilate Convolution and Normal Convolution was used.

In the previous study, 600 images were used for each of 5 classes (person, car, truck, street lamp, tree). However, in this experiment, by increasing the amount of data set, 6 classes (person, car, truck, street lamp, tree, building) and 10,000 images were used each. 17,776 images of tourist spots and 25,825 images of downtown were used, respectively, and a dataset was designed by cropping objects for classification. Train:Test:Val is divided by 8:1:1, and each class consists of the same ratio.

B. Evaluation Metrics

After training for 6 classes with a total of 48,000 images, a learning model was created. And then, we test how many images the learning model can classify correctly. The test dataset has 6,000 images with 1,000 for each class. First, we check the number of images for each class that matches the correct answer. When the accuracy for each class is derived, then the overall accuracy is calculated using the value. Experiment by applying the same method to all comparative models.

C. Implementation Setup

All experiments were conducted at 10 epochs with batch size 64, and the average of accuracy was used as an index after a total of 3 runs. Since the main goal of this thesis is to reduce the weight of artificial neural networks, we also conducted experiments with light weight classification models like squeezenet, shufflenet, mobilenet, etc. Experimental environment was conducted on Intel Core i9-10900X CPU, 4 unit of GeForce RTX 3090s 24GB, and with 188GB RAM memory.

TABLE I
RESULT OF ACCURACY ABOUT EACH CLASSES

Model	Parameters	Accuracy (%)						
		Person	Car	Truck	Street Lamp	Tree	Building	Average
VGG11	132,863,336	99.7	93.4	93.8	**99.4**	99.9	**100**	97.70
ResNet11	11,689,512	99.9	93.8	92.7	99.1	99.8	**100**	97.55
SqueezeNet	1,248,424	99.4	76.7	95.0	98.8	96.5	99.5	94.32
ShuffleNet	1,366,792	98.3	87.4	91.7	98.8	**100**	99.8	97.13
Inception	27,161,264	99.7	89.1	95.4	98.8	**100**	99.8	97.13
GoogleNet	10,332,514	99.7	92.9	**96.1**	99.0	99.8	99.6	97.85
MobileNet	3,504,872	99.6	**94.9**	90.5	99.0	99.8	99.9	97.28
Proposed	**1,032,054**	**100**	94.8	95.1	**99.4**	**100**	99.8	**98.18**

V. RESULT

The proposed network showed 98.18% accuracy performance with the drone image dataset object classification experiment with less number of parameters(1,032,054). Early-stage versions of the classification models like VGG11 [12], ResNet11 [2], Inception [13] and GoogleNet [1] resulted approximately 1% accuracy lower than the proposed method, but the number of parameters was 10 to 100 times more. And representative light weight models like SqueezeNet [14], ShuffleNet [15] and MobileNet [9] also resulted lower than the proposed method at least 0.9% and at most 3.86%. The number of parameters also recorded about 1 to 3 times higher. Trees and buildings with a high proportion of objects in the image had good results. And, unlike other classes, Person and Street Lamp with unique shapes also had high accuracy. However, it showed relatively low values in Car and Truck classification. It is expected that the cause is that both classes can be viewed as Vehicles and are very similar in color and shape. It outperformed good results not only in accuracy but also speed in FPS. All process(train, test, validation) for classifying 60,000 images took 54.43 seconds, which shows 1,102 FPS.

VI. CONCLUSION

This paper conducted a study on the weight reduction of object classification networks using drone image datasets with various characteristics because of taken by many altitude and angle. So, drone image datasets has various features including the front, rear, side and bird-eyes view. To solve this problem, a simple CB and SB were used to classify with high accuracy. The CB use dilate convolution layer because of to watch wide region at once. The SB is the most important part that extracts the features of the input image, and increase the object classification accuracy by putting the output of this block between the networks. In addition, the number of parameters was reduced to about 1,000,000 by reducing the number of channels in the overall network, and fewer parameters were recorded compared to the existing network. The experimental result showed good performance compared to the other networks with an accuracy of 98.18%, and the FPS value also showed good results. In this study, we proved fast and accurate classification work is possible using a simple network with shortcut technique.

REFERENCES

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[2] 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, pp. 770–778.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2016. [Online]. Available: https://arxiv.org/abs/1606.00915

[4] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: https://arxiv.org/abs/1511.07122

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[7] T. Verelst and T. Tuytelaars, "Dynamic convolutions: Exploiting spatial sparsity for faster inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2320–2329.

[8] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[9] 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," *arXiv preprint arXiv:1704.04861*, 2017.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2017.

[11] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1812.01187

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[14] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," *CoRR*, vol. abs/1602.07360, 2016. [Online]. Available: http://arxiv.org/abs/1602.07360

[15] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," 2017. [Online]. Available: https://arxiv.org/abs/1707.01083