Attention Based Object Classification for Drone Imagery

Jehwan Choi
Grad. School of Electrical and Computer Engineering
University of Ulsan, Korea
jhchoi@islab.ulsan.ac.kr

Kanghyun Jo
School of Electrical Eng.
University of Ulsan, Korea
acejo@ulsan.ac.kr

***Abstract*—This paper shows how to make the drone imagery for surveillance or tracking the object in the ground. To detect or classify objects on the ground, convolutional neural networks was adopted and compared with some existed methods and the proposed attention blocks in it. The objects on the ground from the drone images are relatively very small and diversity of the appearance from its perspective projections. This is mainly due to the arbitrary viewpoints from the bird eye views. Furthermore, the distance from its viewpoint in the sky is quite much changeable so that the image of the object is too diverse in appearance and its size. However, the drone is so useful to see widely while navigating in the sky. It is much more attentive to use for real application. Here, some proposed target objects are mainly located in the ground, like static and dynamic objects such as street lamps or trees, vehicles, trucks and pedestrians. These works were done for the national projects to establish the general AI services in Korea recently. For the experiments such as buildup the ground truth of target objects after taken in regulated distance and viewing angles and performed to detect exactly objects in an arbitrary image. For experiments of detection and classification of five categories of objects, attention based CNN architecture was adopted and compared comprehensively with the existed networks like MobileNet, VGG16, SqueezeNet, and ResNet. The experimental results outperformed for the archived drone image dataset with 87.12% in precision. The architecture shows almost 3 times faster with respect to VGG16 or 2 times faster than MobileNet in the speed but a half slimer and twice thicker respectively in the number of parameters. Thus, the Attention Block is useful while a drone navigates through a certain route according to the ground location regardless of the appearance and size of the target region in image.**

***Index Terms*—Attention Block, Computer Vision, Drone Image, Deep Learning, Object Classification**

I. Introduction

For a few decades the object classification in the image has been evolved through the deep learning methodologies, AlexNet [1], VGGNet [2], and GoogleNet [3], and ResNet [4]. Tracking from these ideas were adopted for better detection with deeper layered networks caused the error bound thus the shortcut path ruling out the gradient vanishing and exploding problems. Thus, recent methodolgies reveals much more compounded paths to relieve such caused problems. With this idea, the evolved methods, Xception [5], MobileNet [6], SqueezeNet [7] and others [8]–[10] were developed for increasing accuracy, reducing number of parameters, and the better throughput in computation. This paper proposes to replace the convolution layers with a different size of kernels in a block and apply with different types of convolution methods, like depth-wise or point-wise convolutions. However, the most deep learning networks focus only on improving accuracy, resulting in complex internal structures and a significant number of parameters. This paper adopts a new network to minimize the complexity of network structures and the number of parameters. Also, in this work, the drone imagery was intentionally made and used for detection of the objects in the ground for the autonomous navigation for the future drones. The most common existed object classification dataset consists of views taken from surrounding azimuth viewpoints so that the front, side, and back images were taken at human eye-ball level or so. However, the drones fly above the ground in a different altitude and viewing angles so that the image dataset shows target objects appearance and size are quite arbitrary. However, the number of classes is limited but diverse in appearance. That is, for an object, it can be shown diversely according to the viewpoint. Therefore, it is expected that there will be two major problems in classification. The first is hard to train because of more various characteristics than others. Second, the loss of the characteristic of small images in learning is fatal to learning. However, Attention Block not only adds a variety of features during the learning process, but also re-enters the lost features. That is the reason why using attention block. In addition, we look at what features Attention Block applied networks compare to other networks, and experiment with good performance.

II. Proposed Work

A. Dataset

To create classification dataset, 'Autonomous Drone Flight Video AI Data' is used. These dataset is images of drones flying in the sky and taking pictures of the ground. 'You Only Look Twice' [11] is also taken from the sky, but it takes at an angle of 90 degrees only and the altitude is very high. So there is a difference from the dataset used in this paper. In drone dataset, each image has various altitudes (30 m, 60 m, 100 m, etc.) and angles (15 degrees, 45 degrees, 60 degrees, 90 degrees, etc.). Fig. 1 is one of the images in the dataset. For the data generation, annotating 1,375 images and 3,500 BBox via Yolov5 [12] to construct the object classification dataset. As a result, the class with the most objects is automobiles and 37,738 are extracted, the class with the least objects is traffic

Identify applicable funding agency here. If none, delete this.

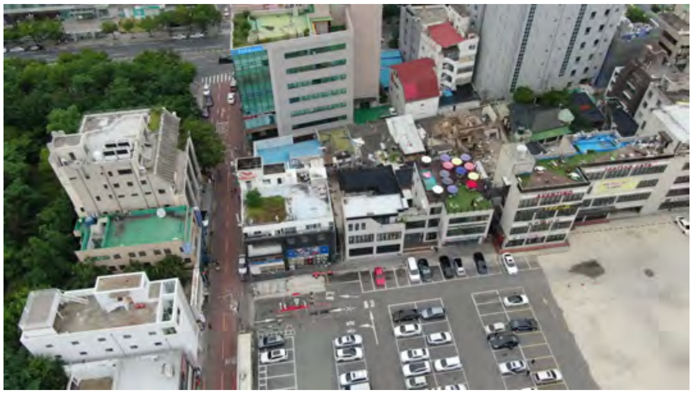

Fig. 1. Example Image of Drone Video Dataset.

signs and 68 are extracted. The CIFAR-10 dataset consists of 6,000 images per class, while the Imagenet dataset consists of 1,000 images per class. Therefore, data is constructed using classes only more than 1,000 objects extracted. More than 1,000 extracted objects are people, car, truck, street lamp, and tree. And there is the number of data per class is 2,000.

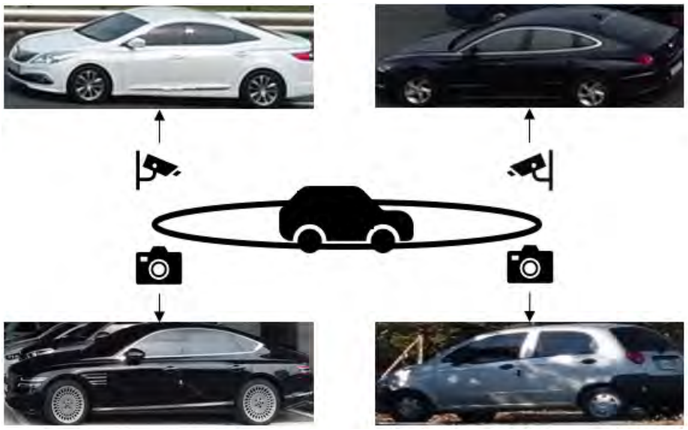

Fig. 2. Existing Dataset Taking Range and Example Images.

Drone dataset has many characteristics. To illustrate the representative characteristics, let me give you an example using car class. In existing datasets, objects are taken by humans or in cctv like Fig. 2. A circle around a car is a path that can be taken based on a car. Therefore, the front, rear, and side of the object are mostly observed, and the top part is also slightly exposed. The size of the image is large and clear because it is taken closely.

On the other hand, when taking with drones, it has a range of semi-spherical shapes based on cars, such as Fig. 3. It would include more angle, especially bird's eye view and plane information than traditional datasets. Thus, the front, rear, side and top of the object are evenly appeared. However, it is different when taken at an angle of 90 degrees. People look like dots and street lamps look like just line because they only see the top. In this case, object classification is difficult, so it is excluded from the dataset configuration. Another characteristic is that the higher the altitude of the taking, the closer the ratio of the horizontal to vertical image is to 1:1. Some objects (people, streetlights) are too small to be recognized by human eyes. The image size of the dataset is 28 × 47, 147 × 432, 257 × 112, etc. The average pixel size of each class is equal to TABLE I, and the average size of the entire image is 120.2 × 146.2. The striking features of the car and the tree are little difference in pixels between horizontal and vertical, making them look more like squares. On the other hand, street lamps are more than twice as long as the length of the vertical. People and trucks differ slightly in length and width. The smallest object being people and the largest being trees. When images are put into the input of the network, different sizes of image are resized to 224 × 224, and train data to 1,500 and test data to 500.

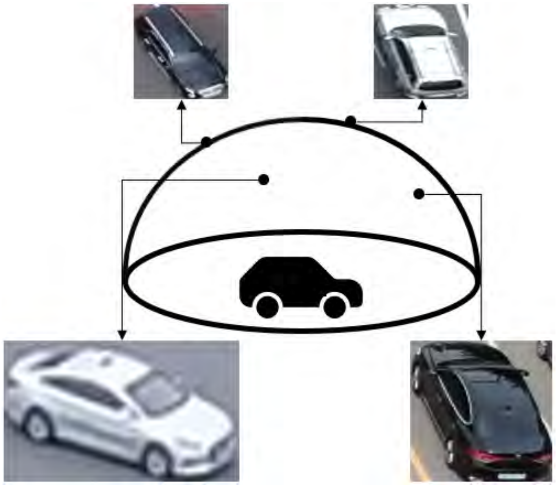

Fig. 3. Drone Dataset Taking Range and Example Images.

TABLE I
AVERAGE PIXEL SIZE OF EACH CLASSES.

Class	Width	Height
Person	30	52
Car	95	89
Truck	146	128
Street Lamp	106	233
Tree	224	229
Average	120.2	146.2

B. Attention Block

In this paper, a network is constructed using attention block technique. Attention block is a proposed technique that complements lost features and extracts more diverse features through convolution layers. A feature map called attention block is add up to the output of the convolution layer before enter to next convolution layer's input. The attention block is applied as shown in Fig. 5.

In Fig. 5, the class on the image is a car and original width and height are about 100 pixels image. The image on the left is the result of resizing to 224 × 224 and inputting it in the

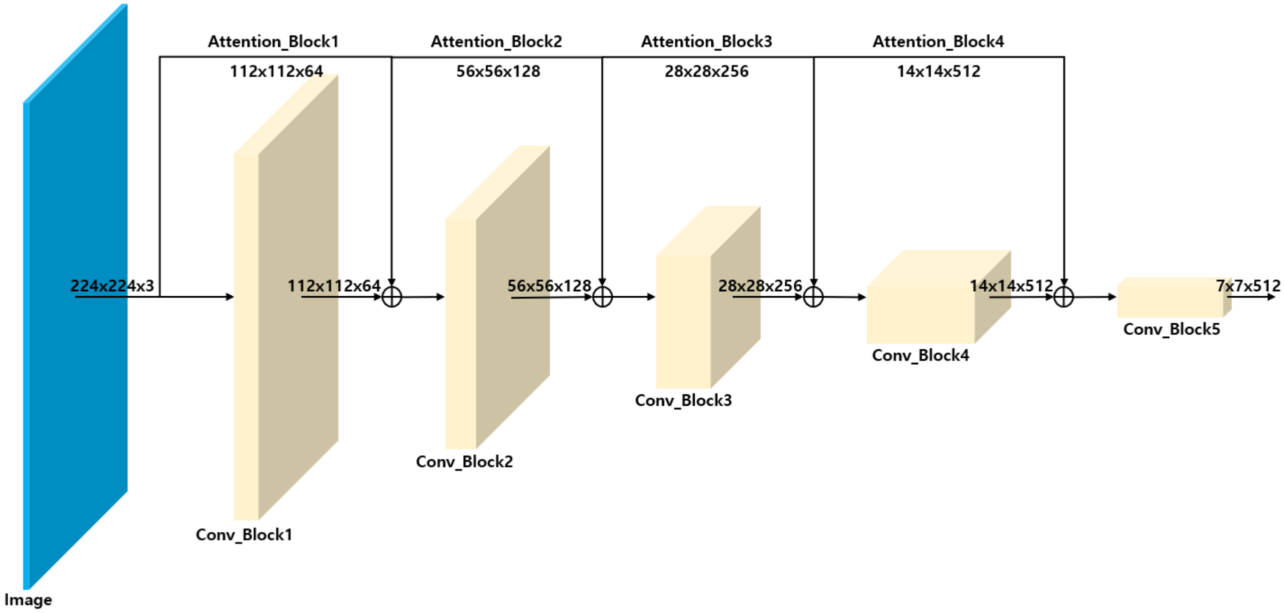

Fig. 4. Proposed Network and Attention Block.

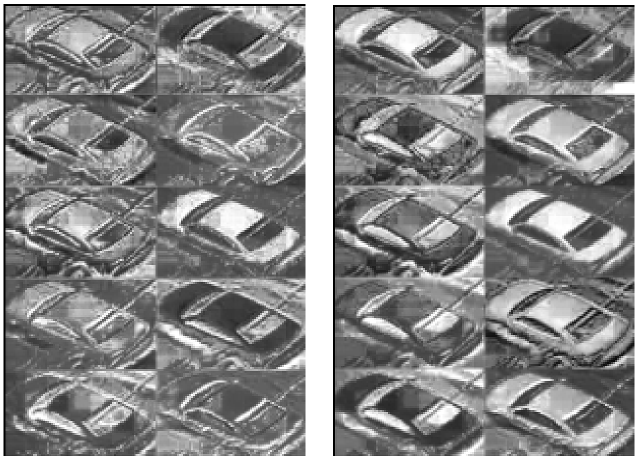

Fig. 5. Output Images after Convolution Layers(left). After Images Applying Attention Map(right).

TABLE II
OVERALL CONFIGURATION OF THE PROPOSED NETWORK.

Layer	Output
Input Image	$224 \times 224 \times 3$
Conv_Block1	$112 \times 112 \times 64$
Attention_Block1	$112 \times 112 \times 64$
Conv_Block1 + Attention_Block1	$112 \times 112 \times 64$
Conv_Block2	$56 \times 56 \times 128$
Attention_Block2	$56 \times 56 \times 128$
Conv_Block2 + Attention_Block2	$56 \times 56 \times 128$
Conv_Block3	$28 \times 28 \times 256$
Attention_Block3	$28 \times 28 \times 256$
Conv_Block3 + Attention_Block3	$28 \times 28 \times 256$
Conv_Block4	$14 \times 14 \times 512$
Attention_Block4	$14 \times 14 \times 512$
Conv_Block4 + Attention_Block4	$14 \times 14 \times 512$
Conv_Block5	$7 \times 7 \times 512$

convolution layer. The right image is the result of adding the attention block to the left image. The effect of applying the attention block is prominent in the image on the right. As such, as the depth of the network deeper, continuing to inject the characteristics of the original image improves the results of the learning.

C. Proposed Method

The networks used in this research consist of convolution block and attention block. The convolution block's internal structure is simple because it consists of a total of four convolution layers and one max-pooling layer. In the convolution layer part, which have two layers with kernel size of 3×3 and two layers with 1×1 using the bottleneck technique. Because of the bottleneck technique maintains similar performance and reduces the number of parameters by approximately 28% is reason of using this method. The convolution block's output will be reduced to quarter size of the input image, and the number of channels will double. The attention block's internal structure is also simple, too. It consists of only 11 convolution layer and max-pooling Layer at first. When convolution block going deeper, each layer of the attention block increases one layer. It only serves to increase the number of channels, reduce the size of the image, to create a new feature map. The reason for configuring as described is to keep the properties of the input image as much as possible and to minimize computations. After these outputs are added, they are passed on to the next convolution layer input and continue to be learned through batch normalization and activation function. The structure of Fig. 4 is described in TABLE II.

III. EXPERIMENT

The experiment environment used Intel Core i9-10900X and four NVIDIA GeForce RTX 2080Ti. Memory used 128GB, and OS used Ubuntu 18.04. The five networks used in this research are all carried out in the same environment, each running separately. The results of experiments with drone flight image object classification dataset are shown in TABLE III. The experimental results adopted the best performance of a total of 100 train and test.

TABLE III
CLASS ACCURACY BY NETWORK USING DRONE FLYING IMAGE OBJECT CLASSIFICATION DATASET.

Class	Proposed	MobileNet	VGG16	SquezzeNet	ResNet
Person	**97.8**	86.8	92.8	96.6	97.0
Car	73.2	84.8	82.4	77.6	**86.0**
Truck	**90.4**	71.6	51.2	67.6	72.0
Street Lamp	80.4	81.8	84.2	**90.0**	75.8
Tree	93.8	92.6	**94.8**	88.8	90.8
Average	**87.12**	85.52	81.08	84.12	84.32

The proposed network in this paper showed the highest accuracy with 87.12% performance in drone image classification dataset experiments. In particular, It performed well in person and truck classes, which are images with an average pixel difference of 10% to 20% between horizontal and vertical. But, car class performed relatively low accuracy. The second best-performing network was the MobileNet [6] with 85.52%.Although there are no outstanding results for one object, the accuracy is high results for all classes. The characteristic of MobileNet [6] is using Depth-wise Separable Convolution, and the ratio of parameters and speed shows good results with relatively very small computations. SqueezeNet [7] and ResNet [4] also showed good results respectively in street lamps and car classes. However, the accuracy was lower than the proposed network. For comprehensive comparison, the image processing speed and size comparison of the network are presented in TABLE V.

TABLE IV
PARAMETERS AND FPS OF NETWORKS.

Model	Parameters	FPS
Proposed	8,251,397	1,538
MobileNet	3,504,872	995
VGG16	138,357,544	555
SqueezeNet	1,248,424	2,061
ResNet	25,557,032	1,024

In TABLE V, SqueezeNet [7] shows the lowest number of parameters and the best FPS. However, the accuracy is lower than others. The proposed network was the second best performance of the five networks, with 1,538 images processed per second and 8,251,397 parameters. Even if the number of parameters more than twice that of MobileNet [6], it can be carculate 500 more images per second.

IV. CONCLUSION

In this paper, experiment to classify objects within drone images was conducted. Drone images used data created by the 'Autonomous Drone Flight Video AI Data' project. Objects in the data have a view from the sky, including human eye level. Therefore, all parts except for the bottom exist in various angles. In other words, it has more characteristics than existing datasets. Thus, object classification experiment was conducted by applying a network using a method called attention block. There are five classes of two types in classification dataset, such as not moved like trees and street lamps, and moved like cars, trucks and persons. The proposed network showed an average accuracy of 87.12%. The results were at least 1.5% and at most 6% higher than the other four networks tested together. The number of parameters was 8,251,397, located in the middle of the comparison group. However, it processed 1,538 images per second and showed a high FPS value compared to the number of parameters. Through this research, the characteristics of drone images were identified, and the use of attention block showed good results for object classification experiments.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

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

[3] 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.

[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, pp. 770–778.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," *arXiv preprint arXiv:1805.09512*, 2018.

[12] D. Thuan, "Evolution of yolo algorithm and yolov5: the state-of-the-art object detection algorithm," 2021.