SETSCI Conference Proceedings
9, 416-419, 2019

4th International Symposium on Innovative Approaches in Engineering and Natural Sciences
November 22-24, 2019, Samsun, Turkey

https://doi.org/10.36287/setsci.4.6.107
2687-5527/ © 2019 The Authors. Published by SETSCI

ISAS WINTER - 2019 ENS

Pedestrian and Vehicles Detection with ResNet in Aerial Images

Enes CENGİZ[1*+], Cemal YILMAZ[1], Hamdi Tolga KAHRAMAN[2] and Fatih BAYRAM[3]

[1]Electrical and Electronics Engineering, Gazi University, Ankara, Turkey
[2]Software Engineering, Karadeniz Technical University, Trabzon, Turkey
[3]Mechatronic Engineering, Afyon Kocatepe University, Afyonkarahisar, Turkey
[*]Corresponding author: enescengiz@aku.edu.tr
[+]Speaker: enescengiz@aku.edu.tr
Presentation/Paper Type: Oral/Full Paper

Abstract – In today's applications, a significant increase in the use of deep learning algorithms is noticeable. The convolution neural network (CNN) of deep learning has been used frequently recently, especially for the successful discrimination of people and vehicles from other objects. Especially with the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, the use of CNNs has become widespread. With the development of technology and traditional image processing techniques, the proses of image processing has been considerably reduced, furthermore, the success rate has increased dramatically. Object detection can be difficult due to the low resolution of objects in aerial images. In this study, a system which automatically recognizes human and different types of vehicles (cars, bicycles, motorcycles) from aerial images taken with drone has been developed. In the system, Residual Networks (ResNet) model, which is the first in the ImageNet competition of the CNN been one of the deep learning techniques, is used. Google Colaboratory with Nvidia Tesla K80 GPU support is used for successful and fast training and testing of the system. In the developed system, results are explained according to different threshold values of the objects detected from the images applied to the input.

Keywords *– Deep Learning, CNN, Image Processing, ResNet, Drone*

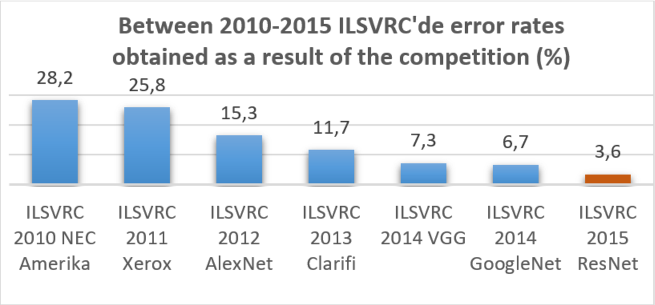
I. Introduction

With the development of technology, drones are widely used in military and civilian applications. They have undertaken many challenging tasks today. Recently drones have been actively involved in many military applications. To ensure the security of a region, it is necessary to have full control of the region. To ensure this security, drones will control the area and prevent any threats. Due to its use in the military field, it needs to be controlled very precisely. Therefore, airborne object detection is emerging as a research area. Detection and tracking of the object to be detected from the air with traditional target detection algorithms will be a problem due to the difficulty of the viewing distance. This problem arises from different regions, noise in surrounding environments and shooting angles. In recent years, with the improvement of data sets and the technological advancement of image processing hardware, Convolution Neural Network (CNN) has made great progress in image recognition and object recognition.

Han et al. are used drone algorithms like Faster RCNN, YOLO etc. for human detection and tracking [1]. Li et al. are worked with the MATLAB program to estimate the number and location of vehicles in a region at the same time [2]. Redmon et al. used YOLO, a connected model for object detection [3].

The inclusion of deep learning algorithms to the ImageNet Large Scale Visual Recognition Competition (ILSVRC) significantly reduced Top-5 errors in 2012. Since then, each year, deep learning models have overcome the challenges and reduced the error rate in the competition.

The biggest leap of deep learning is the 2012 with the CNN, which is accepted as the basic architecture in deep learning [4]. At the 2012 ImageNet competition, Krizhevsky and his colleagues reduced the Top-5 error rate to 15.3% and achieved significant success in this area [5]. The second Top-5 error rate was 26.1%. Since that competition, the Net AlexNet" network has been focused on a wide range of computerized tasks, such as object detection [6], video classification [7], object tracking [8] and super resolution [9]. These achievements created a new field of research that focused on finding high-performance CNN. Since 2014, the quality of network structures has been significantly improved by using deeper and wider. In 2015, the ResNet network reduced the Top-5 error rate to 3.6%, bringing it even below the human error threshold level.



Ross et al. are used the region-based R-CNN framework for target identification [10]. Bayram has proposed a license plate recognition system that enables the identification of the license plate characters from vehicle images with the regional CNN. In order to successfully classify the study, a masked regional CNN based on deep learning is used [11].

In this study, a system for detecting people, cars and bicycles/motorcycles using drone images with convolutional neural network on deep learning which a very popular machine learning approach is presented and results are shown. In the second part, is presented the widely used deep learning algorithm CNN and the ResNet model. In the third part, is presented information about the training and test results of the neural network used in the study. In the fourth part, general results are given about this work.

II. Materials and Method

A. Convolution Neural Network (CNN)

Deep learning is becoming very popular due to innovations in various well-known tasks in the field of artificial intelligence. The reason for using Deep Learning models is that it can answer problems with high accuracy. CNN is a deep learning algorithm that can separate objects in an input image. Its architecture includes Convolution, Pooling, ReLu, DropOut, Fully linked and Classification layers. AlexNet, ZFNet, GoogLeNet, Microsoft RestNet and R-CNN are used in problem solving in Deep Learning [12].

Figure 1 shows the determination of the output by applying the CNN model to the image with a size of 28x28x1 applied to the input. There are certain layers used in neural networks. These; convolution layer, RELU and pooling layers (average and maximum pooling) [13].

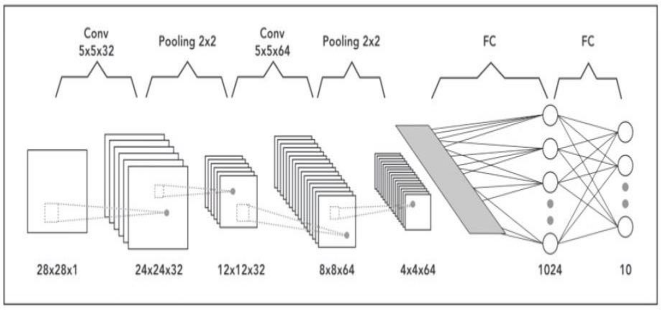


Fig. 1 CNN structure

B. Residual Neural Network (ResNet)

ResNet is a neural network used as a backbone in many computer vision tasks. ImageNet achieved great success by reducing the error rate to 3.6% in the 2015 competition [14]. Because even people depend on their skills and expertise, they have a 5-10% error rate. Figure 2 shows a sequential convolutional network and a ResNet connected convolutional network connection.

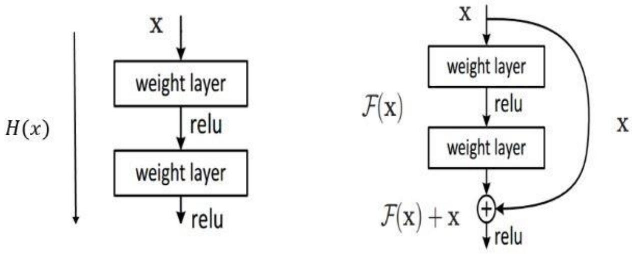


Fig. 2 Sequential convolutional network and ResNet

The greater the number of data sets and layers in deep learning, the more time it takes to form a model. There are some methods to reduce this time. In recent years, some techniques have been proposed for this problem. It uses ImageNet/ResNet to compare training performance in these techniques ([15],[16]). Figure 3 shows the design of a 34-layer ResNet.

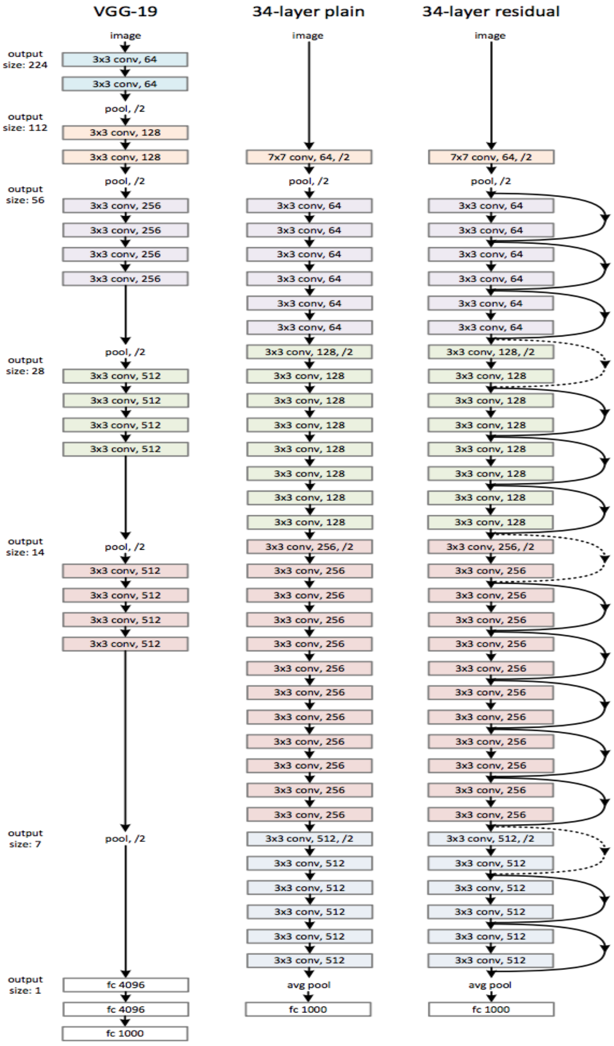


Fig. 3 RestNet architecture 34 layers [14]

C. Experimental Study

In this study, after the identification of the problem in the recognition system, the processes take place step by step. These steps; determining the architecture to be used for deep learning, determining the data set, training and testing ResNet CNN, obtaining the recognition detector, and then evaluating it. The Stanford drone dataset used in the study is a large dataset of aerial imagery collected by the drone over the Stanford university campus. This Dataset is ideal for object detection and tracking issues. Figure 4 shows some examples of images in the drone data set.

Fig. 4 Some examples from the compiled data set for the application

Annotation is required for RetinaNet. All annotations are required according to the format in Equation 1.

$$\text{path/to/image.jpg}, x1, y1, x2, y2, \text{class\_name} \tag{1}$$

The annotation of the images used during training and testing is used in accordance with this format. The system is trained by using 3 different classes as pedestrian, car and bicycle/motorcycle. Table 1 shows the number of images used for training and the total number of annotation objects.

Table 1. Image prepared for training and total number of annotation object

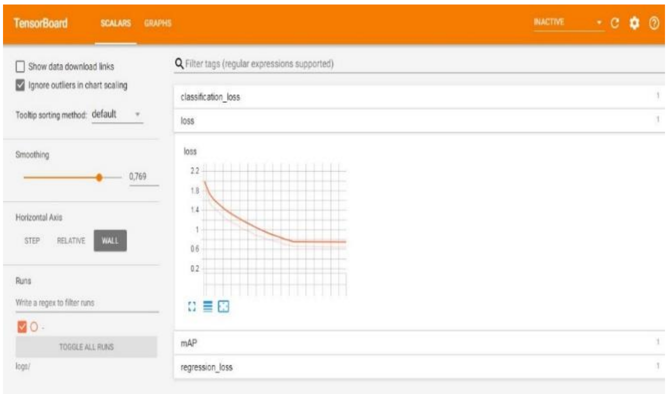
Training Images Number	Total Number of Annotation Objects
2612	35678

Table 2 shows the number of images used for testing and the total number of annotation objects.

Table 2. Image prepared for testing and total number of annotation object

Testing Images Number	Total Number of Annotation Objects
665	6867

Google Colaboratory (Colab) system was used for CNN training and testing. Deep learning activities are developed by using libraries such as Keras, Tensorflow, Pytorch and OpenCV with Colab. Unlike other cloud systems, Colab provides free GPU support. Because the NVIDIA Tesla K80 is a graphics processor in the Colab system, it can present results quickly, as more computations can be made on large data sets. It is expected that losses are decreased in the iterations performed during the training and is fixed while iteration is continuing for a while. Figure 5 shows that initially losses have been reduced and stabilized after a certain iteration.



☐ Show data download links
☑ Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing 0,769

Fig.5 System losses as a result of training at Tensorboard

The input image applied to the model after the training is shown in Figure 6 and the output image taken from the output of our model is shown in Figure 7.

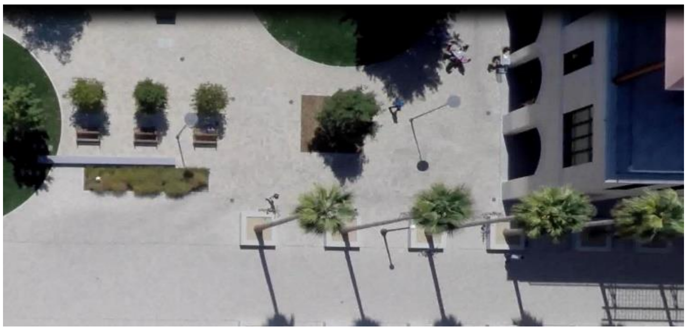


Fig. 6 Image applied to the input of the model

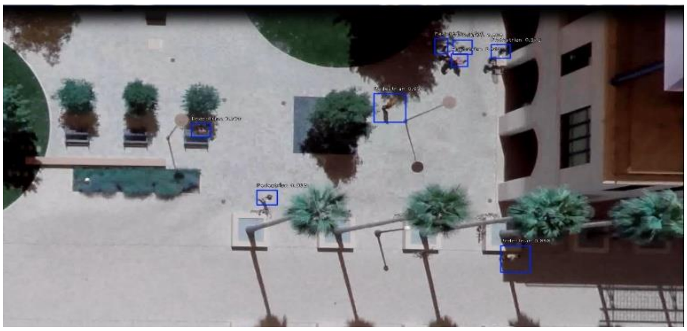


Fig. 7 Image taken to the output of the model

In a study, parameters such as precision, recall and f1 score should be determined to evaluate performance. In this study, precision, recall and F1 score values at various threshold levels are given in Table 3. As can be seen from this table, we have obtained the best F1 score when the threshold value is 0,4.

Table 3: Precision, recall and F1 score values for different threshold levels

Threshold Value	Precision	Recall	F1 Score
0,3	0,73	0,84	0,78
0,4	0,81	0,79	0,80
0,5	0,86	0,73	0,79

III. Results

With drone, it has an extremely high application value in the areas of target detection, targets in sensitive military applications, situation analysis of enemies, personnel search and rescue, agriculture and livestock monitoring. In this study, a system that detects airborne people and vehicles has been developed with deep learning and pre-trained network. Deep learning based convolutional neural network is used to successfully identify people and tools shown in the developed system. For our system 3277 images were used. Approximately 80% of these images were used for training and the remaining 20% were tested in our system. By applying different threshold values, error matrices were extracted. Thus, accuracy, sensitivity and f1 score values were determined. Retina Net detection network based on deep learning has been improved and high performance has been achieved.

REFERENCES

[1] Han, S., Shen, W., & Liu, Z. (2016). Deep Drone: object detection and tracking for smart drones on embedded system.

[2] Li, W., Li, H., Wu, Q., Chen, X., & Ngan, K. N. (2019). Simultaneously Detecting and Counting Dense Vehicles from Drone Images. IEEE Transactions on Industrial Electronics.

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[4] Julia, D. L. f. (2016). "devblogs.nvidia.com." from https://devblogs.nvidia.com/parallelforall/mocha-jl-deeplearning-julia/

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097-1105.

[6] R. Girshick, J. Donahue, T. Darrell and J. Malik. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. (2014). Large-scale video classification with convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 1725–1732. IEEE.

[8] N. Wang and D.-Y. Yeung. (2013). Learning a deep compact image representation for visual tracking. In Advances in Neural Information Processing Systems, 809–817.

[9] Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2), 295-307.

[10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[11] Bayram, F. (2020). Derin Öğrenme Tabanlı Otomatik Plaka Tanıma. Politeknik Dergisi.

[12] Özkan, İ. N. İ. K., & Ülker, E. (2017). Derin Öğrenme ve Görüntü Analizinde Kullanılan Derin Öğrenme Modelleri. Gaziosmanpaşa Bilimsel Araştırma Dergisi, 6(3), 85-104.

[13] Cengil, E., & Çınar, A. (2016). A New Approach for Image Classification: Convolutional Neural Network. European Journal of Technique, 6(2), 96-103.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

[15] Mikami, H., Suganuma, H., Tanaka, Y., & Kageyama, Y. (2018). Imagenet/resnet-50 training in 224 seconds. arXiv preprint arXiv:1811.05233.

[16] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., ... & He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677.