



Düzce University Journal of Science & Technology

Research Article

Classification of Human and Vehicles with The Deep Learning Based on Transfer Learning Method

 Enes CENGİZ ^{a*},  Cemal YILMAZ ^b,  Hamdi Tolga KAHRAMAN ^c

^a Department of Mechatronic Engineering, Faculty of Technology, Afyon Kocatepe University, Afyonkarahisar, TURKEY

^b Department of Electrical and Electronics Engineering, Faculty of Technology, Gazi University, Ankara, TURKEY

^c Department of Software Engineering, Faculty of Technology, Karadeniz Technical University, Trabzon, TURKEY

* Corresponding author's e-mail address: enescengiz@aku.edu.tr

DOI: 10.29130/dubited.842394

ABSTRACT

There has been a significant increase in the use of deep learning algorithms in recent years. Convolutional neural network (CNN), one of the deep learning models, is frequently used in applications to distinguish important objects such as humans and vehicles from other objects, especially in image processing. With the development of image processing hardware, the image processing process is significantly reduced. Thanks to these developments, the performance of studies on deep learning is increasing. In this study, a system based on deep learning has been developed to detect and classify objects (human, car and motorcycle / bicycle) from images captured by drones. Two datasets, the image set of Stanford University and the drone image set created at Afyon Kocatepe University (AKÜ), are used to train and test the deep neural network with the transfer learning method. The precision, recall and f1 score values are evaluated according to the process of determining and classifying human, car and motorcycle / bicycle classes using GoogleNet, VggNet and ResNet50 deep learning algorithms. According to this evaluation result, high performance results are obtained with 0.916 precision, 0.895 recall and 0.906 f1 score value in the ResNet50 model.

Keywords: Deep learning, Object detection, CNN

Derin Öğrenme Tabanlı Transfer Öğrenme Yöntemiyle İnsan ve Araçların Sınıflandırılması

ÖZET

Son yıllarda derin öğrenme algoritmalarının kullanımında önemli bir artış görülmektedir. Uygulamalarda derin öğrenme modellerinden evrişimli sinir ağı (ESA) özellikle görüntü işlemede insan ve araç gibi önemli nesnelere diğer nesnelere ayırmak için sıklıkla kullanılmaktadır. Görüntü işleme donanımlarının gelişmesiyle görüntü işleme süreci önemli ölçüde azaltılmaktadır. Bu gelişmeler sayesinde derin öğrenme üzerine yapılan çalışmaların performansı artmaktadır. Bu çalışmada, dronlar tarafından elde edilen görüntülerden nesnelere (insan, araba ve motosiklet/bisiklet) tespit etmek ve sınıflandırmak için derin öğrenmeye dayalı bir sistem geliştirilmiştir. Derin sinir ağının transfer öğrenme yöntemiyle eğitilmesi ve test edilmesi için açık kaynak olan Stanford Üniversitesi görüntü seti ve Afyon Kocatepe Üniversitesi (AKÜ)'nde oluşturulan drone görüntü seti olmak üzere iki veri seti kullanılmıştır. GoogleNet, VggNet ve ResNet50 derin öğrenme algoritmaları kullanılarak insan, araba ve motosiklet/bisiklet sınıflarını tespit etme ve sınıflandırma işlemine göre kesinlik, duyarlılık ve f1 skor değerleri değerlendirilmiştir. Bu değerlendirme sonucuna göre ResNet50 modelinde 0,916 kesinlik, 0,895 hassasiyet ve 0,906 f1 skor değeriyle performansı yüksek sonuçlar elde edilmiştir.

Anahtar Kelimeler: Derin öğrenme, Nesne tespiti, CNN

I. INTRODUCTION

Target determination has emerged as one of the problems in the field of computer vision, which has attracted the most attention and studies in recent years. The detection and tracking of an object or a living being with the traditional target detection algorithms poses a problem due to the high visibility distance of the object to be detected from the air. In recent years, data driven CNN has been used frequently, with the improvement of datasets and the technological advancement of image processing hardware. In this way, great progress has been made in extracting image information and object recognition [1],[2]. Drones tackle many challenging tasks and contribute to the solution of problems. Drones, which are involved in applications such as object and human detection from the air, search and rescue activities and productivity analysis of agricultural lands, are also encountered in many different areas [3].

Airborne Drone detection of people and vehicles has a very important place for use in the defense industry. It provides great convenience to the user, especially when used in areas that may pose a threat. Determining and classifying the coordinates of the classes in the images is important at this point. The task of detecting and tracking people and vehicles from the images obtained by drone can be overcome with the deep learning algorithms that have recently been presented [4].

How to work the deep learning algorithms in the sub-branch of artificial intelligence can be understood by examining the neural network neurons in the human brain. Deep learning algorithms are prepared by training with a training dataset to solve a problem. The deep learning model, which does not have the desired information at the beginning, learns the necessary information for the problem as a result of the training. Basically, these deep learning models are used in the field of computer vision, which allows the computer to see like a human.

Since the 1990s, important studies have been carried out on object and human classification using deep learning methods. CNN is the best-known model of deep learning. CNNs trained with aerial images obtained from a certain height by means of drones are used in object detection. In the object detection process, the deep learning model is required to show high performance.

In some studies; Vgg network, one of the CNN models, was developed and used for object detection. With the proposed architecture, new layers are added to the network to increase the performance of the model. ResNet connection developed in front of the full connected layers of the Vgg network has been added. It was trained using approximately 20 thousand training and test images to classify 20 different classes and an accuracy rate of more than 85% was achieved [5]. Another study used a deep learning model to detect pedestrians in aerial images. The deep learning models used in the study were trained and tested with the PASCAL VOC 2007 dataset. In order to increase the detection performance of pedestrians in the aerial images included in this dataset, image part and merger transformation was made. Thus, when looking at the experimental results, such preprocessing methods have shown that it can significantly increase the detection rates for CNNs [6]. In the other study, the detection and tracking of pedestrian and car classes was carried out with a deep learning model trained with aerial drone images. Training and testing were carried out separately with two different datasets, one of which is the open source dataset COCO and the other one obtained by drone from the Seoul National University campus. When the results of the study are examined, the model trained with the images obtained from the university campus performed better than the model trained with the open source COCO dataset [7]. In a different study, a deep learning model was proposed for the detection of airborne ships. In the proposed model, the ResNet which is a deep learning model was first trained open source dataset PASCAL VOC 2007. Later, the model was trained by transfer learning with the training set of the SAR ship dataset. The test process of the model was performed with the test set of the SAR ship dataset and the results were presented. The neural network trained with the transfer learning method and the network trained without transfer learning were compared in terms of average sensitivity and training time. According to the comparison result, the network trained with the transfer learning method reveals better performance [8].

In the study, VggNet, GoogleNet and ResNet50 deep learning algorithms are used to detect human, car and motorcycle / bicycle classes. A comparison of algorithms has been made over their success in classifying the specified objects. For the training of the CNNs used, two different datasets were used, one being the open source Stanford drone dataset and the other the drone dataset obtained from the AKU campus we created. Using both datasets, VggNet, GoogleNet and ResNet50 deep network are trained with the transfer learning method and then are tested with the AKU test dataset.

In the first part of this work the importance of aerial human, car, motorcycle / bicycle detection and other previous academic studies on the subject in the literature are mentioned. In the second part, information is given about CNN's working principle. In addition, the layers that make up the CNN are introduced and the process of the network during its training is explained. In the third chapter, the datasets used for training the model, the process of bringing them into the appropriate form and the transfer learning method are explained. Again, in this section, at the end of the test process, by obtaining the precision, recall and f1 score values from the error matrix and is revealed the performance of the model. The fourth section includes the evaluation of the results obtained from the three sections.

II. DEEP LEARNING MODELS

Computer vision is a field that enables machines to see the world as humans, perceive it in a similar way. It is used for many different tasks such as image and video recognition, image analysis and classification, natural language processing. Deep learning is especially used for classification, recognition and detection tasks. The predominant orientation of the studies is in this direction. Nowadays, deep learning algorithms appear in many areas and provide convenience in many tasks based on human power. In recent years, advances in deep learning and computer vision have been built on and perfected on CNN. CNN are multi-layer sensor neural networks used in image recognition and processing, especially designed to process image data [9],[10].

The most useful task to learn how to design a high-performance CNN is to examine high-performance models that have already been built. With the use of CNNs in ILSVRC competition since 2012, successful studies in this field have emerged. In recent years, with the participation of giant technology companies in this competition, its popularity is increasing day by day. ILSVRC competition contributes both to the rapid advancement of the latest technology for computer vision tasks and to the development of general innovations in the architecture of CNN models [4],[11].

The data given as input to CNN is passed through certain layers and transferred towards the output. Convolution can work in 1D (speech processing), 2D (image processing), 3D (video processing). 2D convolution is important because it is used in feature extraction in the field of image processing and is also the basic block of CNN. The image is converted to matrix format by converting the numerical values of each pixel of the colors in the image. The matrix created in this way represents the image numerically. By applying certain filters to this matrix, different details of the entrance image are revealed. That is, it reveals the distinctive features of the image. Some of the filters used are; blur filter, sharpen filter, emboss filter, edge detection filter etc. With the help of these filters, deep attribute information of CNN can be extracted. In deep CNN, which has a multi-layer structure, a separate process is carried out on each layer and data is transferred to the next layer. Each layer performs its own function and finally, the desired classification is made as output [12].

In a neural network, the image applied to the input is passed through layers with multiple neurons and transferred to the output. Each neuron is connected to all neurons in the previous and next layers. There are many layers used in CNN. These; convolution, pooling, activation, flattening are fully connected, normalization and dropout layers [13],[14].

Convolution layer is the process of multiplying filter matrices on the digital matrices of the image in image processing. The pooling layer is the preferred layer after convolution. It is also known as the

down sampling layer [15],[16],[17]. The activation layer is used to regulate the output of layer neurons. There are many types of activation functions, especially the rectified linear unit (ReLU). The task of the flatten layer is to prepare the data at the entrance of the fully connected layer at the end of its network. In this layer, the matrices which comes from the convolution and pooling layer are converted into a one-dimensional array. There are classes to be detected at the exit of the fully connected layer after the flattening layer. The normalization layer in the neural network helps to reduce the training time. This layer resists the vanishing gradient during training of the CNN. Thus, the training time of the network decreases and it shows better performance [18]. The dropout layer is used to prevent it from over-learning the deep neural network. The working performance of the network is increased by preventing from overfitting the network. In this layer, it ensures that a certain percentage of neurons are neglected randomly in each iteration. Dropout takes a value between 0 and 1 [19]. The classification of the object desired from the image applied to the entrance of the deep CNN architecture and the connections between layers are shown in Figure 1.

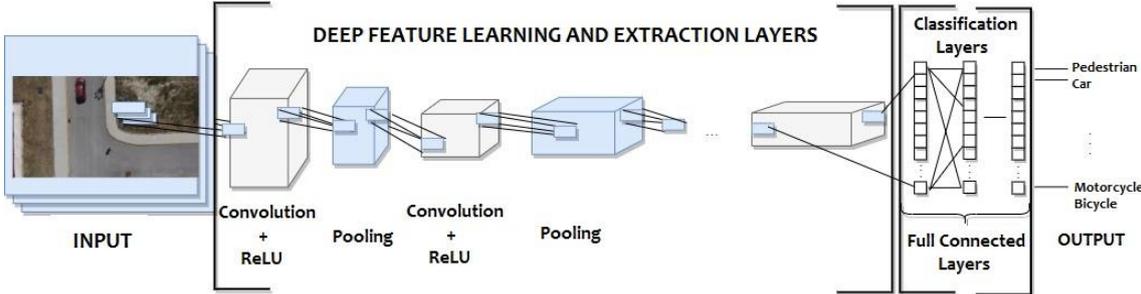


Figure 1. General view of CNN architecture

A project competition called ILSVRC has been organized every year on deep learning since 2010. It is requested that the objects in the images given to the competition participants are detected and classified correctly. CNN architectures have been easily introduced to the world with this competition. With the inclusion of CNN models in the competition in 2012, the emergence of successful models accelerated. The LeNet-5 model is accepted as the first successful and important application of CNN before ILSVRC [20]. Later, CNN models (AlexNet, Vgg, GoogleNet and ResNet) developed for ILSVRC contribute to the literature.

A. 1. AlexNet

The deep learning algorithm AlexNet presented by Alex Krizhevsky et al. Was introduced to the literature in 2012. With the ILSVRC competition, he contributes greatly to the development of CNN. In this model, 224x224 sized images with three color channels are used as input images. The most important difference of the AlexNet model, which is similar to the LeNet architecture, is that it is deeper. The appearance of the AlexNet architecture is shown in Figure 2.

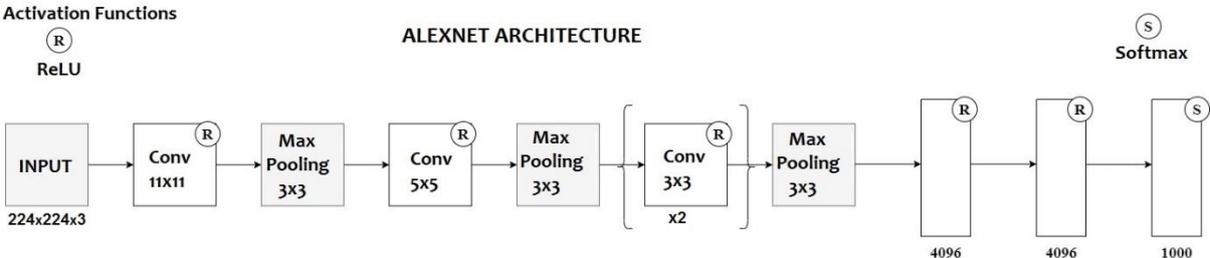


Figure 2. AlexNet CNN architecture

A. 2. VggNet

Vgg architecture is abbreviation of Visual Geometry Group laboratory in Oxford where the studies were conducted. Unlike the large size filters in LeNet-5 and smaller but still relatively large size filters in AlexNet, small size filters (3x3 and 1x1) were preferred in the model. More filters are used compared to other models. The appearance of the VggNet architecture is shown in Figure 3.

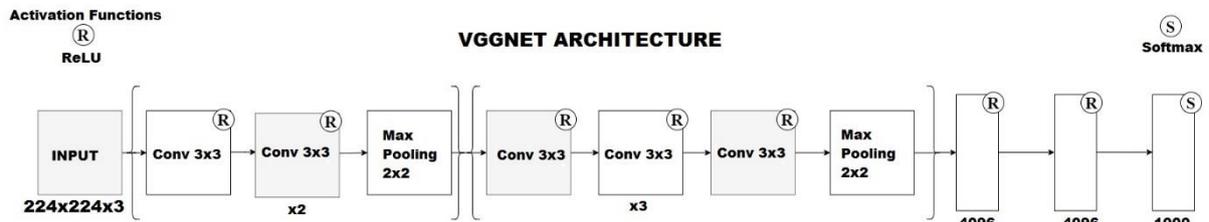


Figure 3. VggNet CNN architecture

A. 3. GoogleNet

Having 12 times less parameters than the AlexNet model, the GoogleNet architecture consists of 22 layers (27 layers when pooling layers are added). This model, also known as the Inception model, was used as the team name of GoogleNet in the ILSVRC14 competition. The architecture, which has a 224x224x3 view at its entrance, has 5 million parameters in total. In the convolution layer, 1x1, 3x3 and 5x5 size filters are used [21]. The filter structure used in the GoogleNet architecture is shown in Figure 4.

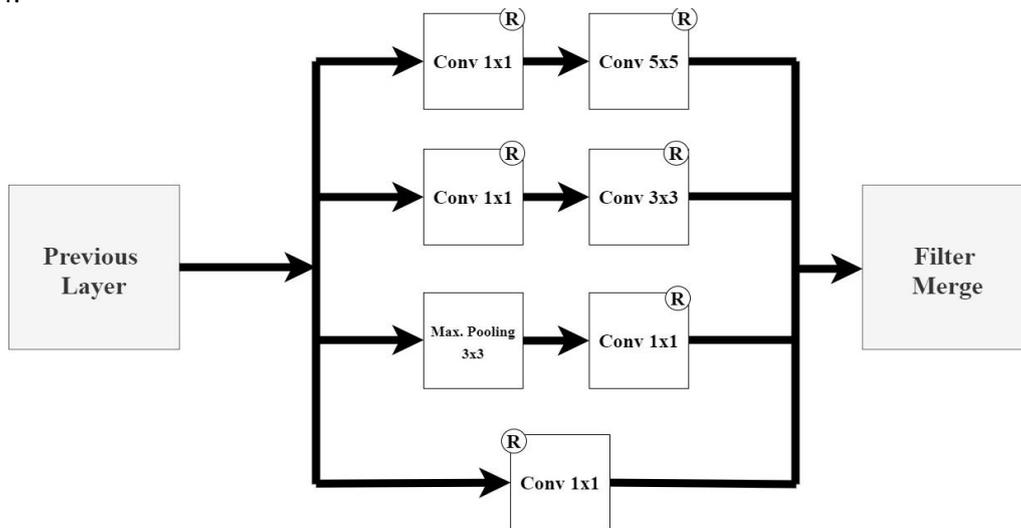


Figure 4. Filter and size reduction structures implemented in GoogleNet architecture

A. 4. ResNet

The ResNet model, known as Residual Network, won the first place in the ILSVRC15 competition, making significant contributions to the literature. ResNet architecture has a deeper structure than other models. With the increase in depth, the target function is approached better and better feature extraction is made. There is a 224x224x3 image at the entrance of the ResNet model. ResNet CNN model generally; It consists of the convolution layer, activation layer, pooling layer and fully connected layer groups [22],[23]. The connection used in the ResNet architecture is shown in Figure 5.

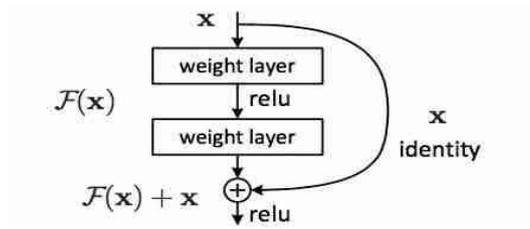


Figure 5. ResNet architecture connection example

Unlike the successive consecutive structure as in the models (AlexNet, VGGNet and GoogleNet) that were developed before it, it used residual blocks in the ResNet model.

III. RELATED WORK

In this section, the application process of the system, which is classified as human, car, motorcycle / bicycle with a deep learning CNN-based model, is explained. The application process of the study is given in the flow chart in Figure 6.

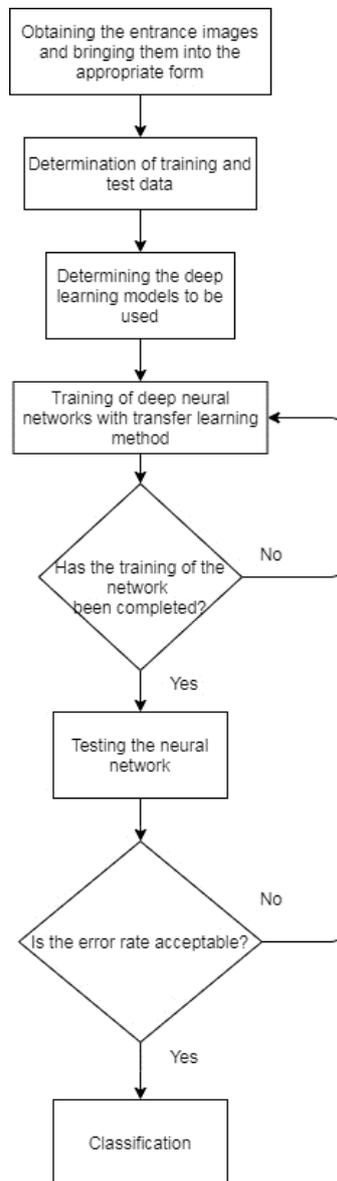


Figure 6. Application process flow chart

Pre-training of the network is provided by training the neural network with the dataset obtained from the Stanford University campus, which is an open source for human, car, motorcycle / bicycle detection and classification [24]. Then, the training of the neural network is continued with the training set of the dataset obtained from the AKU campus. In this way, the performance of the model is increased to higher levels with the transfer learning method. The deep neural network, which is trained with the transfer learning method, is tested with the battery test dataset. The application of the transfer learning method is shown in Figure 7.

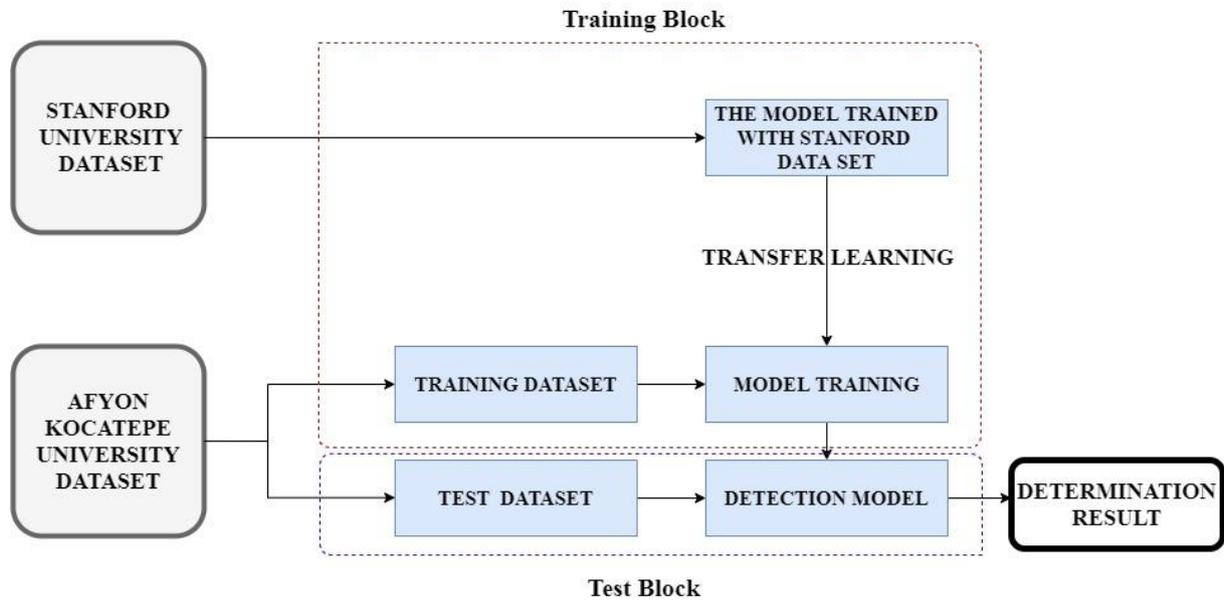


Figure 7. Realization of the transfer learning method to the deep neural network

2591 images are used in the Stanford dataset used for training the deep neural network. These images contain at least one of the objects to be classified. Although the dimensions of the images are 1424x1088 and 1630x1940, they have been reduced to 224x224 for use in deep learning algorithms. Then, drone images are obtained from the AKU campus and the second dataset is created. This dataset consists of approximately 1250 images. These images are 1280x720 in size and have been reduced to 224x224 for training and testing. Training and testing dataset is created by separating 80% for training and 20% for testing of the images in the dataset. The images obtained from the AKU campus are shown in Figure 8.



Figure 8. Images from the created dataset

Weights are updated by training the deep neural network, which is initially randomly weighted, with the Stanford University dataset. Then, the deep neural network is retrained with the AKU training dataset, and the weights are updated. GoogleNet, VggNet and ResNet50 deep learning algorithms are

used for classification of human, car, motorcycle / bicycle. There are certain parameters when training the model in deep learning. Preferred values in these parameters may differ from train to train [25]. Preferred hyper-parameters in the study are given in table 1.

Table 1. Hyper-parameters selected for training

Dataset	Learning rate	Number of training rounds (epoch)	Step Number
Stanford and AKÜ datasets	1e-5	40	2000

In the study, Google Colaboratory system is used to train and test deep learning models. Deep learning studies are developed by using libraries such as Keras, Tensorflow, Pytorch, OpenCV of the system, shortly known as Colab. Thanks to the NVIDIA Tesla K80 graphics processor in the system, more calculations can be made in large datasets, so it can provide results quickly. As the number of iterations increases, the training time of the deep neural network increases. Therefore, training with a GPU supported system results in a much shorter time than a CPU supported system. Examples of the classification process of images are given in Figure 9.



Figure 9. Images Examples of the classification process of images

The above images include the outputs of the images passed from the deep neural network. The outputs are obtained by applying the introduction image to the weights updated with the transfer learning method. Different results are observed with the deep learning models used. The deep learning algorithms using the transfer learning method and their precision, recall and f1 score values are given in Table 2.

Table 2. Precision, recall and f1 score values with transfer learning of deep learning algorithms

Deep Learning Algorithm	Total Image Count	Precision	Recall	F1 Score
GoogleNet	3841	0,882	0,860	0,871
VggNet	3841	0,894	0,872	0,883
ResNet50	3841	0,916	0,895	0,906

The deep learning algorithms without using the transfer learning method and their precision, recall and f1 score values are given in Table 3.

Table 3. Precision, recall and f1 score values without transfer learning of deep learning algorithms

Deep Learning Algorithm	Total Image Count	Precision	Recall	F1 Score
GoogleNet	3841	0,83	0,811	0,82
VggNet	3841	0,839	0,813	0,825
ResNet50	3841	0,854	0,84	0,847

When Table 2 and Table 3 is examined, deep learning algorithms and their precision, recall and f1 score values in images are given. While determining the precision and recall values, the objects in the image are calculated with true positive, false positive, true negative and false negative values. F1 score value gives the harmonic average of precision and recall values.

IV. CONCLUSION

Object detection and classification from images has emerged as a very popular field with the development of deep learning algorithms recently. Effective studies are revealed with the integration of these two developments together with the widespread use of drones in our lives. In this study, a structure that detects and classifies human, car and motorcycle / bicycle classes from the air based on deep learning with drone is created. A new system has been developed by training Stanford drone set, which is open source and AKU dataset we have created with the transfer learning method in deep neural network. In general, high success has been achieved in this detection and classification process performed with 3 different deep learning algorithms. Important studies have been carried out to increase the success of the detection and classification processes of objects such as people and vehicles. Different methods were proposed in each study and had a different effect on success. In this study, models with high performance have been introduced by training the datasets with the transfer learning method. Considering the precision, recall and f1 score values given in Table 2, the ResNet50 model achieved higher success than other models with a score value of 0.906. However, the performance of other models gives higher results compared to models in which transfer learning is not used. When table 2 and table 3 are compared, it is seen that more successful results can be obtained by using the transfer learning method in deep neural networks.

V. REFERENCES

- [1] P. Panchal, G. Prajapati, S. Patel, H. Shah, and J. Nasriwala, "A review on object detection and tracking methods, " International Journal for Research in Emerging Science and Technology, vol. 2, no. 1, pp. 7-12, 2015.
- [2] H. Li, Z. Wu, and J. Zhang, "Pedestrian detection based on deep learning model, " In 2016 9th International Congress on Image and Signal Processing, Bio Medical Engineering and Informatics, pp. 796-800, 2016.
- [3] M. Hassanalian, and A. Abdelkefi, "Classifications, applications, and design challenges of drones: A review," Progress in Aerospace Sciences, vol. 91, pp. 99-131, 2017.
- [4] U. Shah, and A. Harpale, "A Review of Deep Learning Models for Computer Vision," In 2018 IEEE Punecon, Pune, India, pp. 1-6, 2018.
- [5] M. F. Haque, H. Y. Lim, and D. S. Kang, "Object Detection Based on VGG with ResNet Network," In 2019 International Conference on Electronics, Information, and Communication, Auckland, New Zealand, pp. 1-3, 2019.
- [6] Y. C. Chang, H. T. Chen, J. H. Chuang, and I. C. Liao, "Pedestrian Detection in Aerial Images Using Vanishing Point Transformation and Deep Learning," In 2018 25th IEEE International Conference on Image Processing, Athens, Greece, pp. 1917-1921, 2018.
- [7] H. Song, I. K. Choi, M.S. Ko, J. Bae, S. Kwak, and J. Yoo, "Vulnerable pedestrian detection and tracking using deep learning," In 2018 International Conference on Electronics, Information, and Communication, Honolulu, USA, pp. 1-2, 2018.

- [8] Y. Li, Z. Ding, C. Zhang, Y. Wang, and J. Chen, "SAR Ship Detection Based on Resnet and Transfer Learning," In IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, pp. 1188-1191, 2019.
- [9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.
- [10] L. Deng, and D. Yu, "Deep learning: methods and applications, " Foundations and Trends in Signal Processing, vol. 7, no. 4, pp. 197-387, 2014.
- [11] C. Kyrkou, G. Plastiras, T. Theocharides, S. I. Venieris, and C. S. Bouganis, "DroNet: Efficient convolutional neural network detector for real-time UAV applications, " In 2018 Design, Automation & Test in Europe Conference & Exhibition, Dresden, Germany, pp. 967-972, 2018.
- [12] K. K. Çevik, and A. Çakı, "Görüntü İşleme Yöntemleriyle Araç Plakalarının Tanınarak Kapı Kontrolünün Gerçekleştirilmesi," *Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, c.. 10, s. 1, ss. 31-38, 2010.
- [13] F. Bayram, "Derin öğrenme tabanlı otomatik plaka tanıma, " *Politeknik Dergisi*, c. 23 , s. 4, ss. 955 - 960, 2020.
- [14] A. Kızırcak, and B. Bolat, "Derin Öğrenme ile Kalabalık Analizi Üzerine Detaylı Bir Araştırma, " *Bilişim Teknolojileri Dergisi*, vol.11, pp. 263-286, 2018.
- [15] E. Cengil, and A. Çınar, "A New Approach for Image Classification: Convolutional Neural Network, " *European Journal of Technic*, vol. 6, no. 2, pp. 96-103, 2016.
- [16] W. Rawat, and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review, " *Neural computation*, vol. 29, no. 9, pp. 2352-2449, 2017.
- [17] T. Pala, U. Güvenç, H. T. Kahraman, İ. Yücedağ, and Y. Sönmez, "Comparison of Pooling Methods for Handwritten Digit Recognition Problem, " In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-5, 2018.
- [18] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," In International conference on machine learning, pp. 448-456, 2015.
- [19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting, " *The Journal of Machine Learning Research*, vol. 15, no. 1, 1929-1958, 2014.
- [20] Y. LeCun, L. D. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, and V. Vapnik, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural Networks: The Statistical Mechanics Perspective*, New Jersey, USA, 261-276, 1995.
- [21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions, " In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.
- [22] 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, Las Vegas, USA, pp. 770-778, 2016.

- [23] E. Cengiz, C. Yılmaz, H. T. Kahraman, and F. Bayram, "Pedestrian and Vehicles Detection with ResNet in Aerial Images," 4th. International Symposium on Innovative Approaches in Engineering and Natural Sciences, Samsun, Turkey, pp. 416-419, 2019.
- [24] T. Tang, Z. Deng, S. Zhou, L. Lei, and H. Zou, "Fast vehicle detection in UAV images," In 2017 International Workshop on Remote Sensing with Intelligent Processing, Shanghai, China, pp. 1-5, 2017.
- [25] J. H. Yoo, H. I. Yoon, H. G. Kim, H. S. Yoon, and S. S. Han, "Optimization of Hyperparameter for CNN Model using Genetic Algorithm," In 2019 1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), pp. 1-6, 2019.