COMPARATIVE STUDY OF DEEP LEARNING MODELS IN MULTI-LABEL SCENE CLASSIFICATION

Saziye Ozge Atik[1]
[1]Gebze Technical University, Cayirova Campus, Turkey
Email: soatik@gtu.edu.tr

KEY WORDS: Multi-class classification, Scene classification, Deep learning

ABSTRACT: Many state-of-the-art studies are being conducted on environmental monitoring with computer vision applications. Remotely sensed images are widely preferred data in this regard. The body of your abstract begins here. Many open data sets have been generated in this area, and multi-class classification is carried out automatically with the studies carried out. Behalf of human interpretation, using machine learning algorithms provides economic, time, and robust utilities. UC Merced Land Use dataset is one of the most common datasets, including a wide variety of classes in the meaning of land use. In the study, seven different deep learning models are conducted to the UC Merced Land use dataset, and multi-class land use classification results have been compared quantitatively. The algorithms yielded higher than %95 accuracies. The highest overall accuracy was obtained using the DenseNet 121 model, and the worst score was obtained with Alexnet. In several test images, using the SqueezeNet model provided more successful predictions for several classes. In future studies, domain-shift applications can strengthen the studies for more expansive areas.

1. INTRODUCTION

Remotely sensed data is one of the primary sources for computer vision tasks. Computer vision applications for earth observations have various studies such as aerial scene classification (Zheng *et al.*, 2019; Bi *et al.*, 2020; Dede *et al.*, 2018) object detection (Atik and Ipbuker, 2020; Sezen *et al.*, 2022; Cepni *et al.*, 2020), semantic segmentation [Duran *et al.*, 2021; Atik *et al.*, 2021; Atik and Ipbuker, 2021). Scene classification can be grouped under three main groups, as shown in Figure 1. The binary classification has two classes, and images can refer to only one of them. In multi-class classification, there are many classes, and images are labeled as one of the labels. Therefore, the images can have more than one label for multi-label classification, and the class number is more than two. Datasets containing aerial scenes can be used for multi-class or multi-label classification. Additionally, classification type depends on the scope of the dataset and category type.

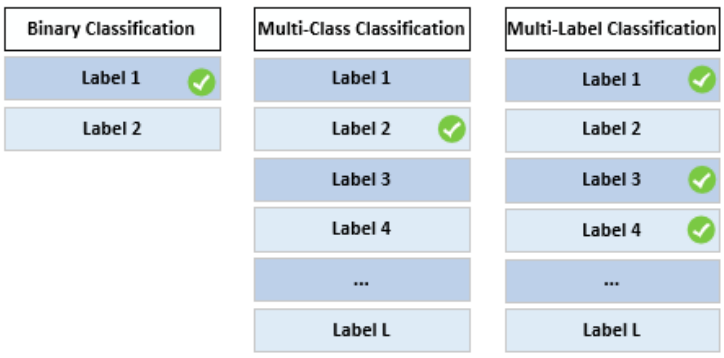

Figure 1. Binary, Multi-Class and Multi-Label Classification

In this study, the application of multi-label classification with different Convolutional Neural Networks (CNNs) was carried out. UC Merced Land Use (Yang and Newsam, 2010), a high-resolution aerial images data set, is used for this purpose.

2. DATA AND METHODOLOGY

The UC Merced dataset contains 100 images (256 x 256) for each class belonging to 21 classes. The sample images of the class are shown in Figure 2. In general, the dataset includes many artificial structures and natural classes. For example, the residential class is split into three classes in the UCM dataset.

Figure 2. Images of UC Merced Data Set Classes (The image is adapted from (Özyurt et al., 2020)).

The aerial images in the dataset also contain more than one class. Therefore, the total number of classes in the dataset is various. The total number of labels of images is shown in Figure 3 per class. Several multi-label samples are shown in Figure 4.

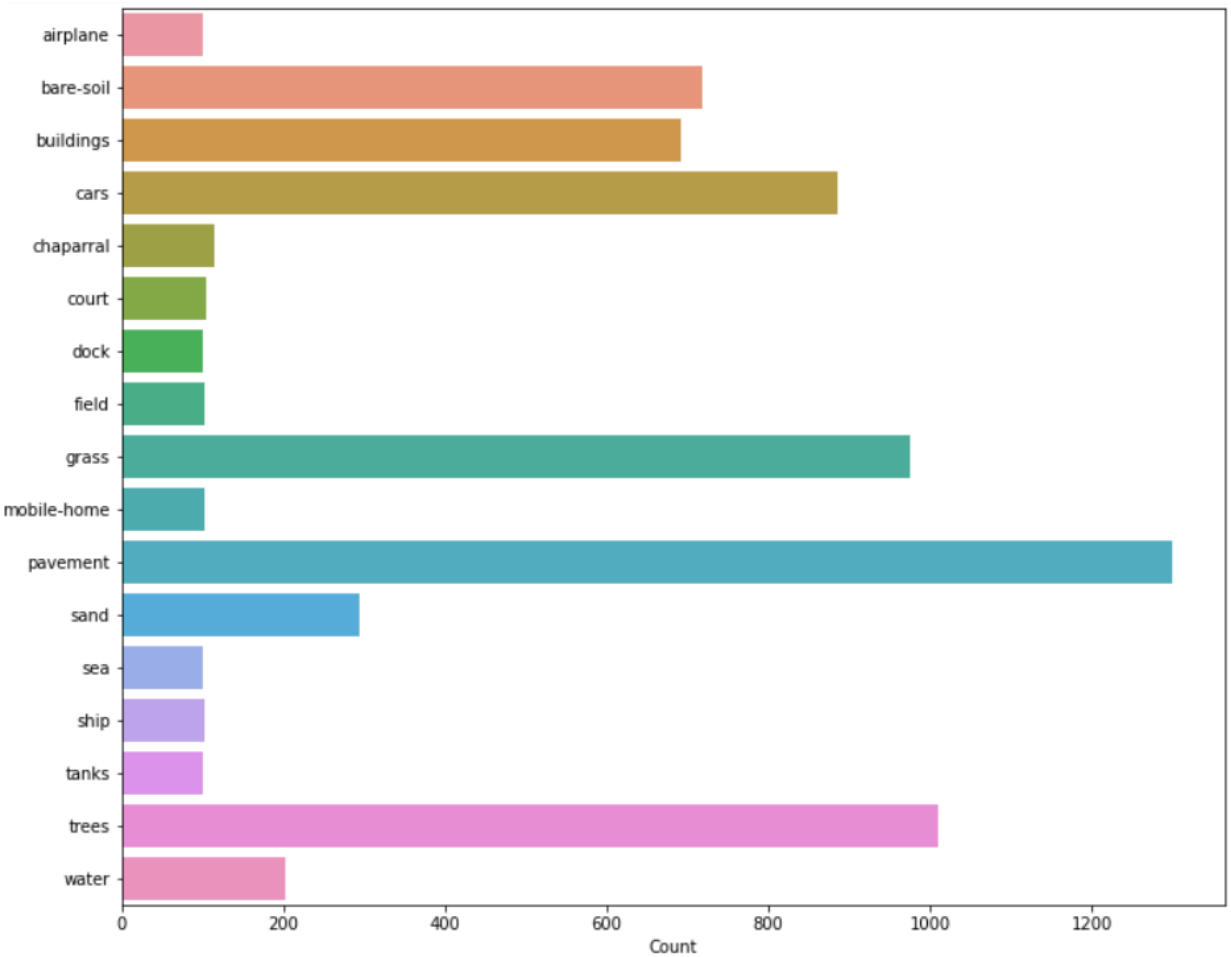

Figure 3. UC Merced data set class counts as multi-label classification

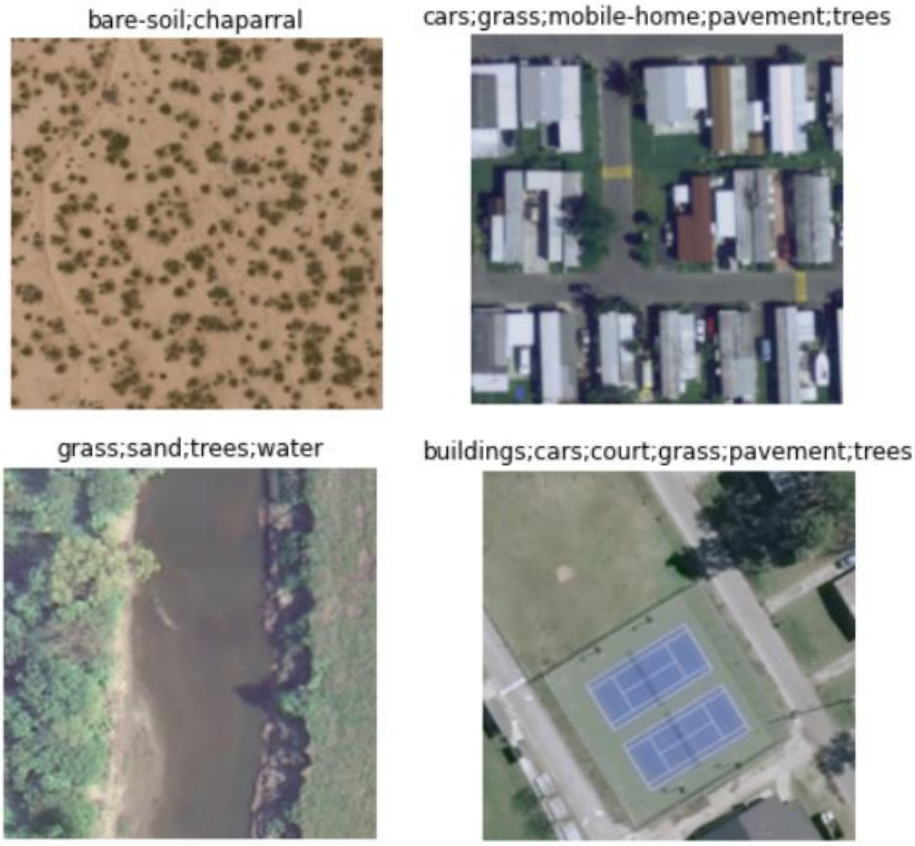

Figure 4. Multi-label classification samples of UC Merced Land Use data set

2.1 AlexNet

Krizhevsky *et al.* (2012) trained DCNN 1.2million high-resolution images for 1000 different classes in the ImageNet LSVRC-2010 contest as AlexNet.

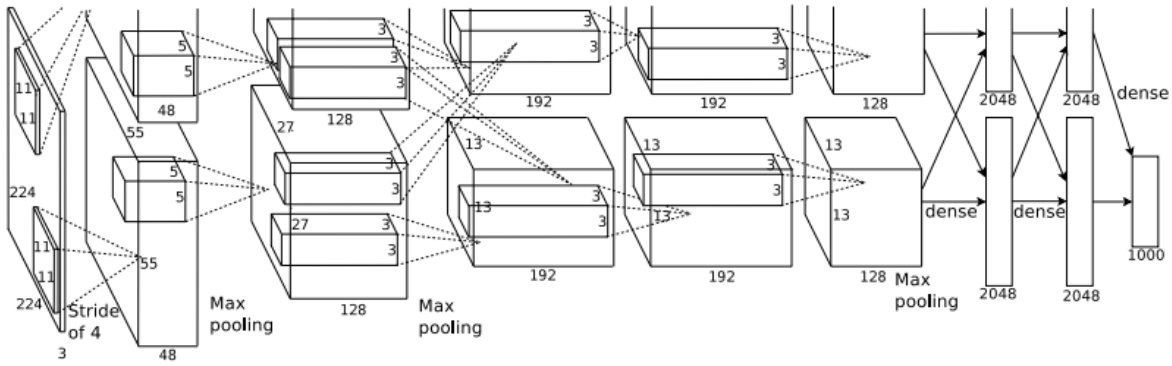

Figure 5. An illustration of AlexNet architecture [12]

The architecture has a local normalization scheme that aids generalization (Equation 1). Here $a^i_{x,y}$ is the neuron's activity and kernel *i* is at the position (x, y). Applying ReLU (Rectified Linear Unit) nonlinearity and $b^i_{x,y}$ is the response-normalized activity. In the expression, k=2, n=5, α=10, and β=0.75.

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \sum_{\min(N-1, i+\frac{n}{2})}^{\min(N-1, i+\frac{n}{2})} (a^j_{x,y})^2 \right)^{\beta} \quad (1)$$

The effectiveness of the architecture is also verified on the CIFAR-10 dataset.

2.2 ResNet

In Figure 4, the residual learning scheme is shown for ResNet architecture. H(x)I the formula is F(x)=H(x)-x used in the deep residual learning framework.

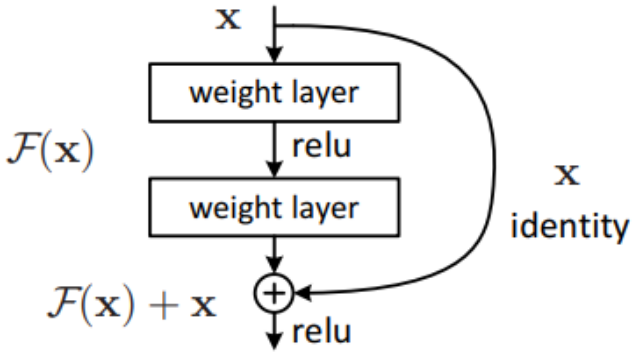

Figure 4. Residual learning: a building block [13].

In the formulation of $\mathcal{F}(\times) + \times$ is a feedforward neural network that has shortcut connections.
$y = \mathcal{F}, \{W_{i\}}) + \times$ defines the building block, and the dimension of x and $\mathcal{F}$ must be equal. x and y refer to the input and output vectors of the layers. F(x, {Wi}) refers to multiple convolutional layers (He *et al.*, 2016).

2.3 DenseNet

Huang et al. generated a CNN as Dense Convolutional Network (DenseNet). It connects each layer to other layers with a feedforward type (Figure 5). The network has several advantages, and they evaluate it in four competitive trend benchmarks: CIFA-10, CIFAR-100, SVHN, and ImageNet (Huang *et al.*,2017)

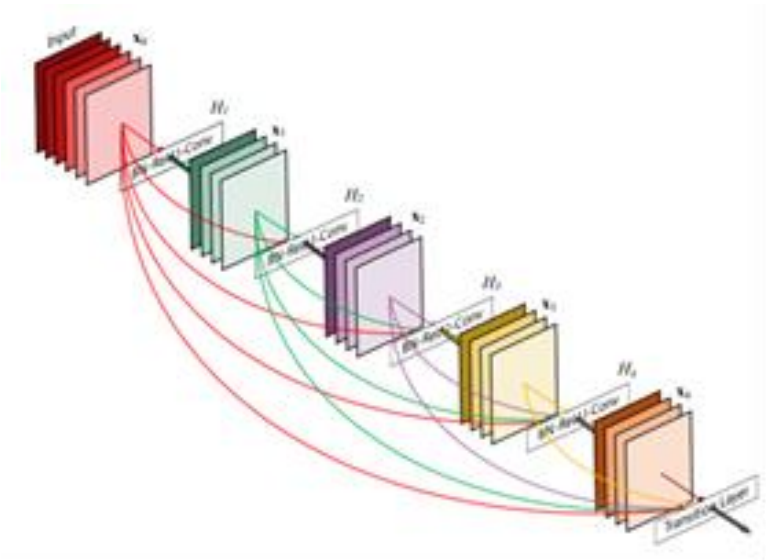

Figure 5. An illustration of the 5-layer dense block with a growth rate of k = 4 [14].

2.4 SqueezeNet

Iandola *et al.* (2016) proposed a method that provides AlexNet level accuracy on ImageNet using fifty times fewer parameters. In addition, the network has several advantages about being smaller CNNs with less communication phase at training, requiring diminished bandwidth, and being proper even with less capacity of memory [15]. In the literature, there are other versions, such as SqueezeSeg (Wu *et al.*, 2018)for semantic point cloud segmentation tasks (Atik and Duran, 2022).

3. RESULT AND DISCUSSION

Pre-trained DL models were used in the training phase. The epoch number was selected as ten, determined experimentally, and it is the same for all CNN models. Error rate and time are observed for each model. 30% of the datasets were split randomly as test data. After the testing phase, the network results are compared with evaluation metrics for the classes by mean precision, recall, and F1 score. In Equation 2, overall accuracy is explained. The equation includes TP: true positive, FN: false negative, FP: false positive, and N: total classification number.

$$Overall\ Accuracy = \frac{TN+TP}{N} \tag{2}$$

SquuezeNet and ResNet models have been used in the experiments as different versions. In addition, AlexNet and DenseNet CNN models are used for multi-label aerial scene classification of the UC Merced Land Use data set.

Table 1. Overall Accuracy of CNN Models

Model	Overall Accuracy
AlexNet	**95,18**
SquuezeNet 1_1	95,73
SquuezeNet 1_0	96,01
ResNet18	96,18
ResNet34	96,39
ResNet50	96,77
DenseNet121	**96,89**

The highest performance has been obtained with DenseNet121 with %96,89 overall accuracy. Moreover, the worst overall accuracy has been obtained with AlexNet with %95.18. The overall accuracy of CNN models is shown in Table 1 quantitatively. Also, the accuracy comparison is shown in Figure 6.

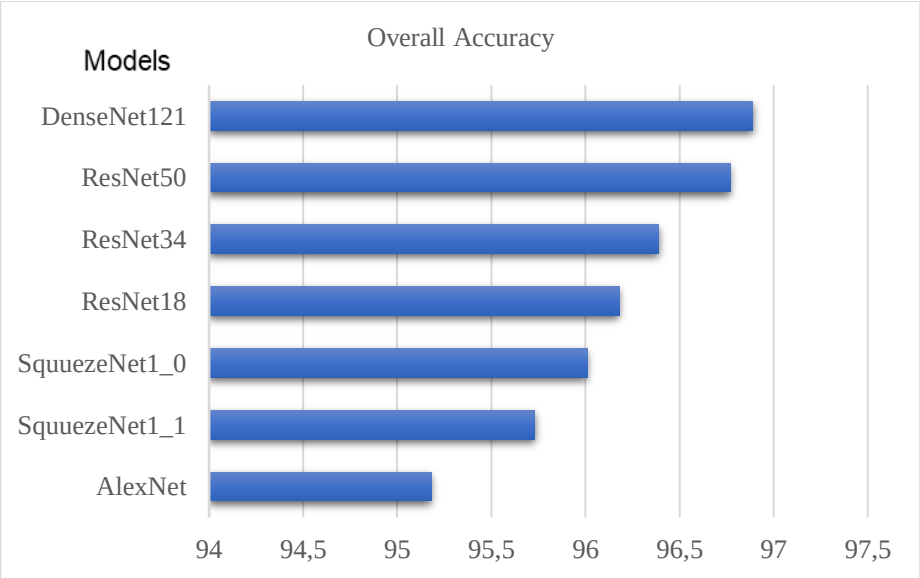

Figure 6. Overall Accuracy Graph of the CNN Models

In future projections, other backbone models of these CNNs can be used for these data sets for multi-label classification purposes. Also, domain-shift applications can be performed while training in one data set and testing in another data source. These studies help enhance the Sustainable Development Goals (SDGs) of the United Nations (UN) for understanding aerial images through artificial intelligent approaches.

REFERENCES

Atik, M. E., & Duran, Z. (2022). An Efficient Ensemble Deep Learning Approach for Semantic Point Cloud Segmentation Based on 3D Geometric Features and Range Images. Sensors, 22(16), 6210

Atik, M. E., Duran, Z., & Seker, D. Z. (2021). Machine learning-based supervised classification of point clouds using multiscale geometric features. ISPRS International Journal of Geo-Information, 10(3), 187.

Atik, S. O., & Ipbuker, C. (2020). Instance Segmentation Of Crowd Detection In The Camera Images. In Proceeding of Asian Conference on Remote Sensing.

Atik, S. O., & Ipbuker, C. (2021). Integrating convolutional neural network and multiresolution segmentation for land cover and land use mapping using satellite imagery. Applied Sciences, 11(12), 5551.

Atik, S. O., Atik, M. E., & Ipbuker, C. (2022). Comparative research on different backbone architectures of DeepLabV3+ for building segmentation. Journal of Applied Remote Sensing, 16(2), 024510.

Bi, Q., Qin, K., Li, Z., Zhang, H., Xu, K., & Xia, G. S. (2020). A multiple-instance densely-connected ConvNet for aerial scene classification. IEEE Transactions on Image Processing, 29, 4911-4926.

Cepni, S., Atik, M. E., & Duran, Z. (2020). Vehicle detection using different deep learning algorithms from image sequence. Baltic Journal of Modern Computing, 8(2), 347-358.

Dede, M. A., Aptoula, E., & Genc, Y. (2018). Deep network ensembles for aerial scene classification. IEEE Geoscience and Remote Sensing Letters, 16(5), 732-735.

Duran, Z., Ozcan, K., & Atik, M. E. (2021). Classification of Photogrammetric and Airborne LiDAR Point Clouds Using Machine Learning Algorithms. Drones, 5(4), 104.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. Squeezenet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv 2016, arXiv:1602.07360.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Özyurt, F., Ava, E., & Sert, E. (2020). UC-merced image classification with CNN feature reduction using wavelet entropy optimized with genetic algorithm.

Sezen, G., Cakir, M., Atik, M. E., & Duran, Z. (2022). DEEP LEARNING-BASED DOOR AND WINDOW DETECTION FROM BUILDING FAÇADE. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 43, 315-320.

Wu, B., Wan, A., Yue, X., & Keutzer, K. (2018, May). Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1887-1893). IEEE.

Yang, Y., Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10). Association for Computing Machinery, New York, NY, USA, pp. S270–S279.

Zheng, X., Yuan, Y., & Lu, X. (2019). A deep scene representation for aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing, 57(7), 4799-4809.