ISSN: 2148-9173 Vol: 7 Issue:3 Dec 2020

International Journal of Environment and Geoinformatics (IJEGEO) is an international, multidisciplinary, peer reviewed, open access journal.

Comparison of Fully Convolutional Networks (FCN) and U-Net for Road Segmentation from High Resolution Imageries

Ozan OZTURK., Batuhan SARITURK, Dursun Zafer ŞEKER

Abstracting and Indexing: TR DIZIN, DOAJ, Index Copernicus, OAJI, Scientific Indexing Services, International Scientific Indexing, Journal Factor, Google Scholar, Ulrich's Periodicals Directory, WorldCat, DRJI, ResearchBib, SOBIAD

IJEGEO *International Journal of Environment and Geoinformatics 7(3):272-279 (2020)*

Reaserch Article

Comparison of Fully Convolutional Networks (FCN) and U-Net for Road Segmentation from High Resolution Imageries

Ozan Ozturk*, Batuhan Sariturk, Dursun Zafer Şeker

ITU, Department of Geomatics Engineering, Faculty of Civil Engineering, Istanbul, TR

* Corresponding author: Ozan Ozturk
* E-mail: oozturk16@itu.com

Received 15 May 2020
Accepted 15 Sep 2020

How to cite: Ozturk, et al., (2020). Comparison of Fully Convolutional Networks (FCN) and U-Net for Road Segmentation from High Resolution Imageries. *International Journal of Environment and Geoinformatics (IJEGEO)*, 7(3):272-279 . DOI: 10.30897/ijegeo.737993

Abstract

Segmentation is one of the most popular classification techniques which still have semantic labels. In this context, the segmentation of different objects such as cars, airplanes, ships, and buildings that are independent of background and objects such as land use and vegetation classes, which are difficult to discriminate from the background is considered. However, in image segmentation studies, various difficulties such as shadow, image blockage, a disorder of background, lighting, shading that cause fundamental modifications in the appearance of features are often encountered. With the development of technology, obtaining high spatial resolution satellite imageries and aerial photographs contain detailed texture information have been facilitated easily. Parallel to these improvements, deep learning architectures have widely been used to solved several computer vision tasks with an increasing level of difficulty. Thus, the regional characteristics, artificial and natural objects, can be perceived and interpreted precisely. In this study, two different subset data that were produced from a great open-source labeled image sets were used to segmentation of roads. The used labeled data set consists of 150 satellite images of size 1500 x 1500 pixels at a 1.2 m resolution, which was not efficient for training. In order to avoid any problem, the imageries were divided into smaller dimensions. Selected images from the data set divided into small patches of 256 x 256 pixels and 512 x 512 pixels to train the system, and comparisons between them were carried out. To train the system using these datasets, two different artificial neural network architectures U-Net and Fully Convolutional Networks (FCN), which are used for object segmentation on high-resolution images, were selected. When the test data with the same size as the training data set were analyzed, approximately 97% extraction accuracy was obtained from high-resolution imageries trained by FCN in 512 x 512 dimensions.

Keywords: Deep Learning, Image Segmentation, Fully Convolutional Networks (FCN), U-Net

Introduction

Image segmentation has become a significant topic of interest in the remote sensing field due to the ever-increasing quantity of high spatial resolution (HSR) imagery acquired from satellites, airplanes, unmanned aerial vehicles (UAVs), and other platforms. Image segmentation can be formulated as a classification problem of pixels with semantic labels. Segmentation commonly used for medical image analysis, autonomous vehicles, video surveillance, and augmented reality, etc. For this context, image segmentation is one of the most fundamental, useful, and popular topics in image processing and analysis methods (Minaee et al., 2020).

Compared with other image processing approach, image segmentation is one of the most challenging tasks. This method aims to change the representation of an image into something more meaningful and easier to analyze. The result of image segmentation is a set of segments that collectively cover the entire image. In segmentation, some parameters of each pixel, such as color, intensity, and texture (Barghout et al., 2003).

Image segmentation and object detection in optical remote sensing images generally suffer from several increasing challenges such as view variations, shadow, and occlusion. Early studies, the low spatial resolution of earlier satellite images (such as Landsat), would not allow the detection of separate human-made or natural objects. These studies generally focused on extracting the region properties from these images. With the advances of remote sensing technology, the very high resolution (VHR) satellite (e.g., IKONOS, SPOT-5, and Quickbird) and aerial images that generally generated unmanned aerial vehicles (UAV) have been providing more detailed spatial and textural information. Because of the increased resolution, a greater range of human-made objects can be separately identified (Cheng and Han, 2016).

Object detection and segmentation methods can be divided into four main groups which are; matching-based methods, knowledge-based methods, machine learning-based methods, and deep learning-based methods (Esetlili et al., 2018; Cheng and Han, 2016). In the past, numerous image segmentation algorithms have been developed in the literature, from the earliest methods, such as thresholding, histogram-based bundling, region growing, k-means clustering, watersheds, to more advanced algorithms such as active contours, graph cuts, conditional and Markov random fields, and sparsity-based methods (Minaee et al., 2020). Contrary to the machine learning-based methods, others are not suitable for complex processing, which generally end with mislearn. Parallel to the development of machine learning methods, learning-based approach have begun to increase their effectiveness in these kinds of tasks. Neutral networks made massive progress because a large amount of data is available, and the computing power which is getting faster as GPUs, etc. become general-purpose computing tools. During the last decades, works in deep learning dealing with image segmentation have been significantly improved by using neural networks.

Considerable efforts have been made to develop various methods for the image segmentation of different types of objects in satellite and aerial images such as roads, cars, buildings, roofs, and airplanes. In these methods, rules are prepared by the processing of data and answers such as features and labels of these features. Moreover, layered representations learning and hierarchical representations of data can be carried out in deep learning. Numerous image segmentation algorithms have been developed, such as U-Net (Ronneberger et al., 2015), Fully Convolutional Neural Network (Long et al., 2015), ParseNet (Liu et al., 2015) and SegNet (Badrinarayanan et al., 2017). These techniques based on the simulation of the learning process for decision and no need to write complex programs. More recently, deep learning-based algorithms have been dominating the top accuracy benchmarks for the various image segmentation task such as biomedical image segmentation (Ronneberger et al., 2015), Automatic brain tumor segmentation (Wang et al., 2017), mobile vision application (Howard et al., 2017) and building segmentation (Cheng et al., 2019). During the last decades, deep learning (DL) networks have yielded a new generation of image segmentation models with remarkable performance improvements resulting in what many regards as a paradigm shift in the field with achieving the highest accuracy rates on popular benchmarks (Minaee et al., 2020).

In this study, deep learning-based image segmentation was used. Two different artificial neural network architectures of U-Net and Fully Convolutional Networks (FCN) were compared for road segmentation. Additionally, in the study, together with the challenges of current studies, promising research directions in future studies were discussed.

Deep Learning

Machine learning, deep learning, and artificial intelligence have been the subject of lots of different studies, such as intelligent cities, meteorological estimates, and change detection analysis. In Figure 1, the relationship among artificial intelligence, machine learning, and deep learning (Çelik and Gazioğlu, 2020).

Artificial intelligence was born in 1950, emerged with the question of whether computers can solve the problems solved by people. The main goal is to automate the activities performed by people. In this context, artificial intelligence encompasses machine learning and deep learning. Early studies in artificial intelligence, due to manipulating the knowledge, programmers believed that a clear and broad set of rules should be created. This approach is known as symbolic artificial intelligence. Popularity increased with the rise of expert systems. For instance, the chess that was one of the first applications of artificial intelligence, was able to solve logical problems quickly and accurately. However, it was insufficient to solve fuzzy logic problems such as image classification, speech recognition, and language translation (Chollet, 2018).

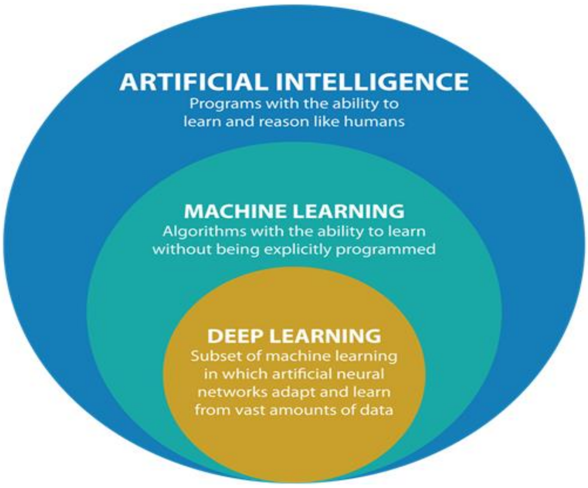

ARTIFICIAL INTELLIGENCE
Programs with the ability to learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn without being explicitly programmed
DEEP LEARNING
Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Figure 1. A categorization of artificial intelligence (Digitalogy, 2019)

Machine learning consists of questions such as whether the computer can learn to perform a specific task when processing the data, whether the rules can be learned automatically. In classical programming, rules and data are processed, and results are produced. However, in machine learning, rules are obtained by processing data and answers. A machine-learning system is trained rather than explicitly programmed. It is most important to train the networks for determining the model and rules in the machine learning approach. For instance, to automate traffic flow density, a machine-learning system would learn statistical rules for associating specific pictures with many different data such as weather, changing lighting, and traffic conditions.

Deep learning is a specific subset area of machine learning, and it examines computer algorithms that learn and improve on their own. The term "deep" in deep learning refers to the number of layers in the model. Inspired by the nerve cells that make up the human brain, neural networks comprise layers (neurons) that are connected in adjacent layers to each other like a human brain. In other words, deep learning techniques is represented the number of layers which contribute to a model of the data. The main goal of deep leaning is layer-based representations of learning and visualize the hierarchical representations (Chollet, 2018). The network learns something simple at the initial level in the hierarchy and then sends this information to the next level. The next level takes this simple information, combines it into something a bit more complicated, and passes it on the last level. This process continues as each level in the hierarchy builds something more complex from the input it received from the previous level.

Deep learning has been achieved successful results comparing other approaches such as near human-level image classification, digital assistants and ability to answer natural-language questions, etc. Deep learning opened a new era in many fields, from image classification to image segmentation. Generally, deep learning techniques can be divided into five categories according to the basic method, which is; Convolutional Neural Networks (CNNs), Recurrent Neural Network, Restricted Boltzmann Machines (RBMs), Autoencoder and Sparse Coding. The categorization of deep learning methods, along with some representative works, is given in Figure 2.

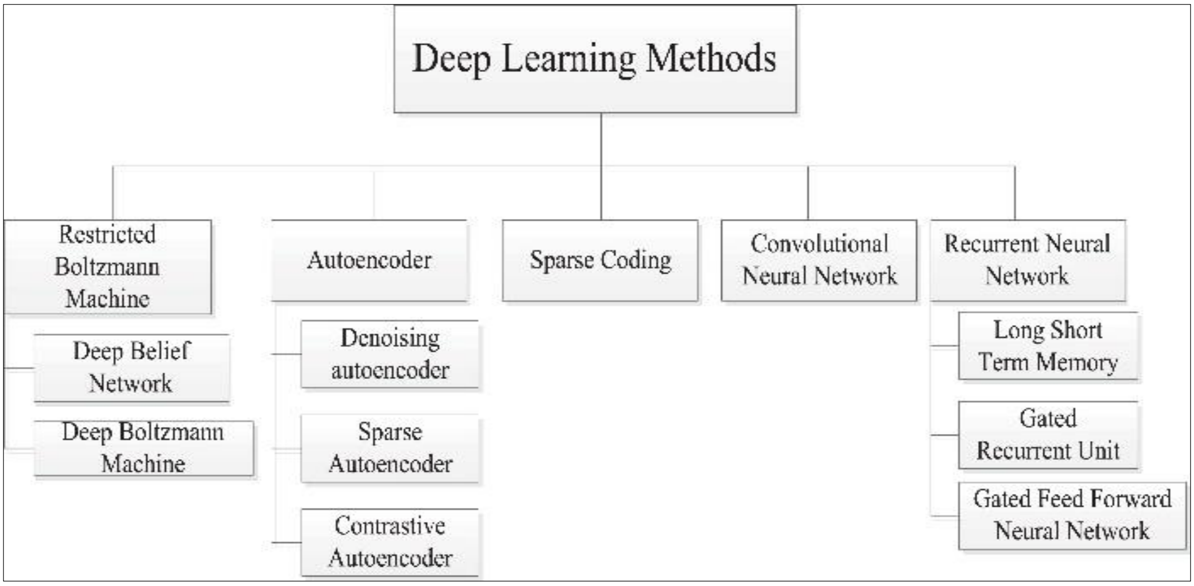


Figure 2. A categorization of the deep learning methods (Nweke et al., 2018).

Different kinds of layers show different roles. Deep learning has been widely adopted in various tasks of computer vision, such as image classification, object detection, object recognition, and image segmentation. Image segmentation is the most popular and promising task for deep learning. Successful results obtained from CNN models which are capable of tackling the pixel-level predictions with the pre-trained networks on large-scale datasets. Convolutional Neural Networks (CNN) is the most commonly used computer vision applications in image processing. Multiple layers are trained robustly on CNN. Generally, CNN consists of three main neural layers, which are convolutional layers, pooling layers, and fully connected layers (Guo et al., 2016).

CNN's image segmentation with CNN involves feeding segments of an image as input to a convolutional neural network, which labels the pixels. CNN cannot process the whole image at once. It scans the image, looking at a small "filter" of several pixels each time until it has mapped the entire image. The researchers modified existing CNN architectures, such as VGG16 and GoogLeNet, to manage non-fixed sized input and output by replacing all fully-connected layers with the fully-convolutional layers. As a result, the model outputs a spatial segmentation map instead of classification scores. Afterward, there are several models initially developed for medical/biomedical image segmentation, which are inspired by FCNs and encoder-decoder models. U-Net is a well-known, such architectures, which are now also being used outside the medical domain (Minaee et al., 2020; Erdem and Avdan 2020; Ronneberger et al., 2015).

Materials and Methods

In this study, comparisons were made between two CNN architectures, U-Net and Fully Convolutional Networks (FCN), using two different datasets in sizes of 256 x 256 and 512 x 512 to extract roads from high-resolution images.

The dataset used in the study is Toronto Roads Dataset, created by Volodymyr Mnih and distributed by Toronto University. The Toronto Roads dataset consists of roughly 500 square kilometers of training data, 48 square kilometers of test data, and 8 kilometers of validation data at a resolution of 1.2 m. This dataset contains both urban and suburban areas of Toronto. Toronto Roads dataset contains some omitted roads with minor registration problems. In figure 3, examples of this data set have been presented.

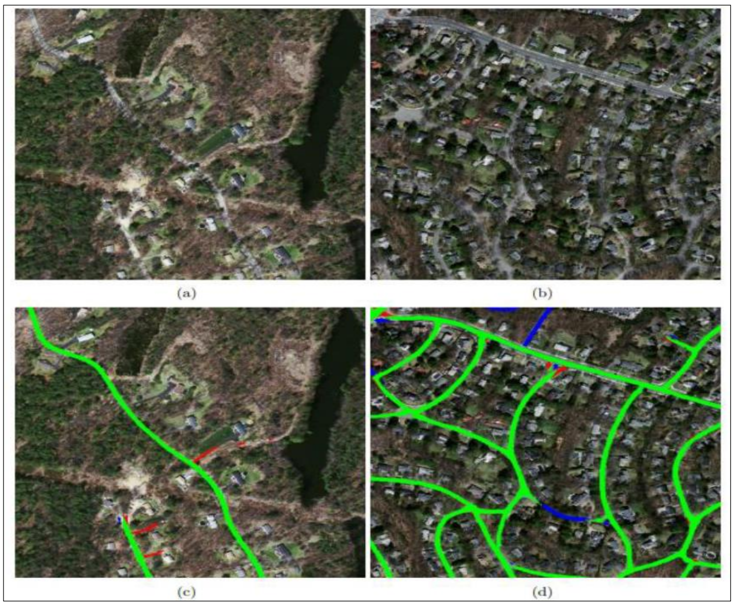


Figure 3. a and b are aerial images from different areas, c and d are labeled images from these areas.

To train the models, images, and their corresponding labels were divided into smaller patches of 256 x 256 and 512 x 512 pixels to increase the number of samples, reduce the computational cost, analyze the effects of image size on the model and not to lose resolution with resizing operation.

Fully Convolutional Neural Networks (FCNN) based segmentation, replacing the fully connected layers with more convolutional layers and it has been a popular strategy and baseline for semantic segmentation (Chen et al., 2014). Chen et al., 2014 proposed a similar FCN model but also integrated the strength of conditional random fields (CRFs) into FCN for detailed boundary recovery. Long et al., 2015 defined architecture that combined semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. The workflow of the FCNN has been presented in Figure 4.

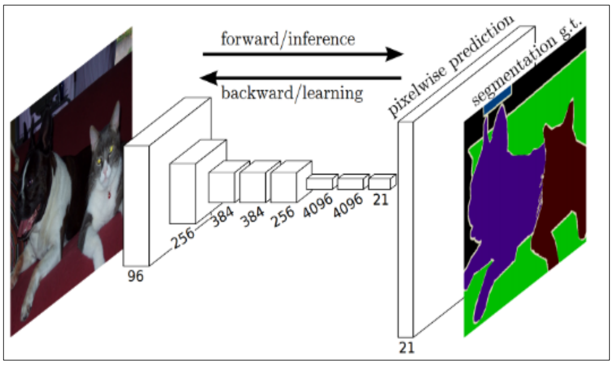


Figure 4. FCN workflow diagram (Long et al., 2015)

Each layer of data in a Convnet is a three-dimensional array of size h × w × d, where h and w are spatial dimensions, and d is the feature or channel dimension. The first layer is the image, with pixel size h × w, and d color channels. Locations in higher layers correspond to the locations in the image they are path-connected to, which are called their receptive fields. Convnets are built on translation invariance. Their basic components (convolution, pooling, and activation functions) operate on local input regions and depend only on relative spatial coordinates. An FCN operates on an input of any size and produces an output of corresponding (possibly resampled) spatial dimensions. Moreover, when receptive fields overlap significantly, both feedforward computation and backpropagation are much more efficient when computed layer-by-layer over an entire image instead of independently patch-by-patch.

Another model, U-Net is a convolutional network architecture use for fast and precise segmentation of images. It was developed at the Computer Science Department of the University of Freiburg (Ronneberger et al., 2015). Up until now, it has outperformed the prior best method (a sliding-window convolutional network) on the ISBI challenge for the segmentation of neuronal structures in electron microscopic stacks. It won the Grand Challenge for computer-automated Detection of Caries in Bitewing Radiography at ISBI 2015, and it won the Cell Tracking Challenge at ISBI 2015 on the two most challenging transmitted light microscopy categories. U-Net workflow is presented in Figure 5.

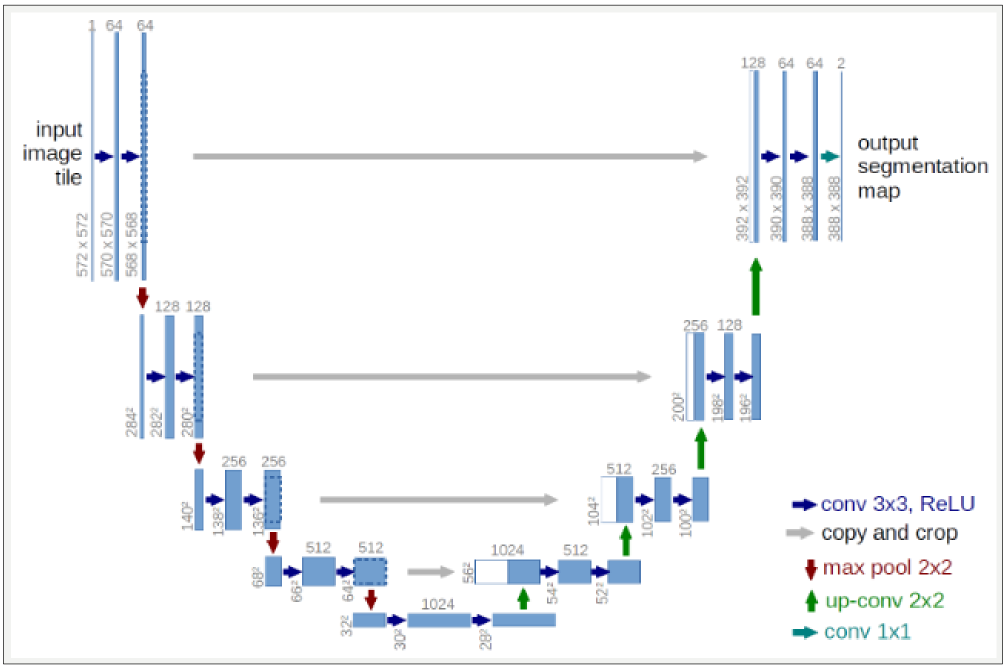


Figure 5. U-Net workflow diagram (Ronneberger et al., 2015)

In this figure, each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower-left edge of the box. White boxes represent copied feature maps. The arrows indicate the different operations. This architecture is based on Long et al. 2015's Fully Convolutional Network. It was modified to work with fewer training features. U-Net's main differences compared to FCN are; U-Net is symmetric, and it uses skip connections between downsampling and upsampling paths.

Training the Model

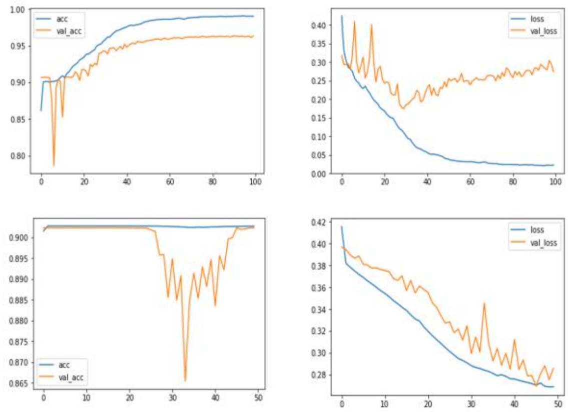
Original U-Net and FCN architectures were used to train models using both datasets. From each dataset, randomly selected 1200 samples were used for training, while the remaining 300 samples were used for validation. Used optimizers and hyperparameters for training are listed in Table 1.

Table 1. Used hyperparameters

	U-Net (256)	FCN (256)	U-Net (512)	FCN (512)
Optimizer	Adam	Adam	Adam	SGD
Learning Rate	0.0001	0.0001	0.0001	0.01
Batch Size	16	16	16	4
Epochs	100	100	100	80

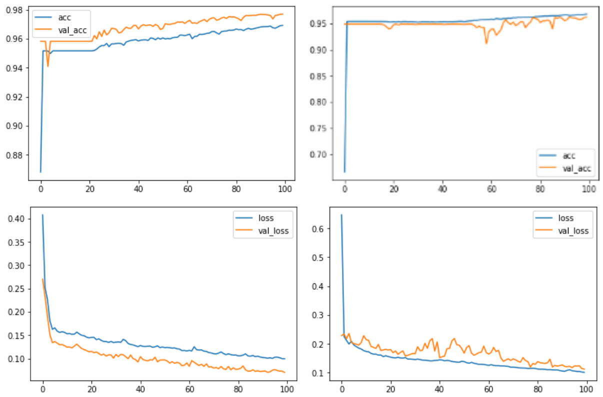
Accuracy and loss assessment results of 256 x 256 U-Net and 256 x 256 FCN architectures are shown in Figure 6, and the results of 512 x 512 architectures are shown in figure 7.

In graphics, blue curves indicate training accuracy, and orange curves show validation loss rate.



FCN U-Net

Figure 6. Accuracy and loss assessment results of U-Net architecture and FCN (256 x 256)



FCN U-Net

Figure 7. Accuracy and loss assessment results of U-Net architecture and FCN (512 x 512)

Results

The validation accuracies of U-Net and FCN architectures for 256 x 256 datasets have been calculated as 96.33% and 90.23%, respectively. The accuracy of U-Net architecture is 6.1% higher than FCN architecture for the 256 x 256 dataset. For dataset 512 x 512, the validation accuracy of U-Net architecture calculated as 96.18%, and the accuracy of FCN architecture calculated as 97.69%. The accuracy of FCN architecture is 1.51% higher than U-Net architecture for 512 x 512 dataset.

Sample segmented images obtained by the models are shown in Figures 8 and Figure 9. In predicted images, yellow pixels show segmented roads and purple pixels show background.

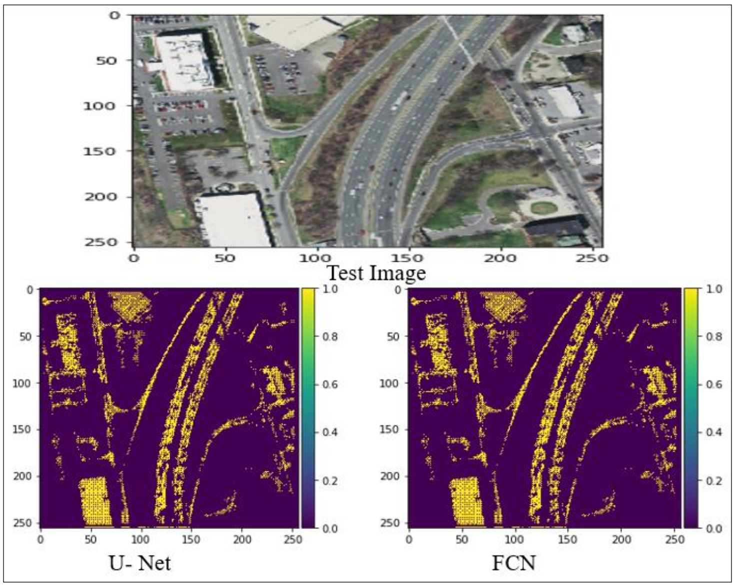


Figure 8. Segmentation results from test image (256 x 256)

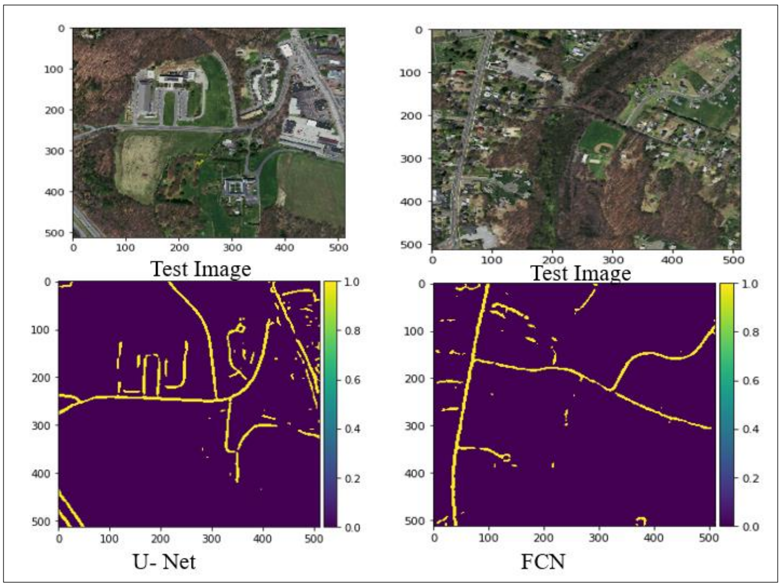


Figure 9**.** Segmentation results from test image (512 x 512).

The final accuracy results are shown in Figure 10. When training and validation accuracies of the models were compared, it can be seen that the FCN model using a 256 x 256 dataset has the lowest accuracy for both training and validation. Additionally, overfitting was observed for the FCN model that uses the same dataset. For models that use a 512 x 512 dataset, training accuracies are similar, but the FCN model is more accurate than the U-Net model by 1.5% according to validation accuracy results. When datasets compared, it can be seen that 512 x 512 sized datasets performed better than 256 x 256 sized datasets.

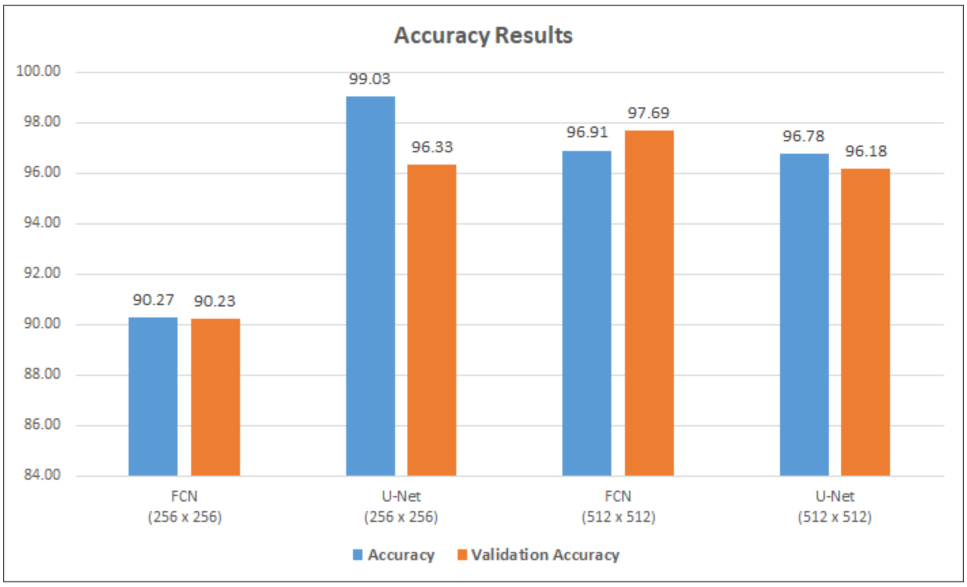


Figure 10. Training and validation accuracy results of the models.

The final loss results are shown in figure 11. When training and validation losses of the models were compared, models that use a 512 x 512 sized dataset outperformed the models that use a 256 x 256 sized dataset. Similar to accuracy results, U-Net performed better with 256 x 256 dataset, and FCN performed better with 512 x 512 dataset according to loss results. Similarly to the accuracy results, overfitting was observed with the U-Net model that uses a 256 x 256 sized dataset, as can be seen from the difference between training and validation loss results.

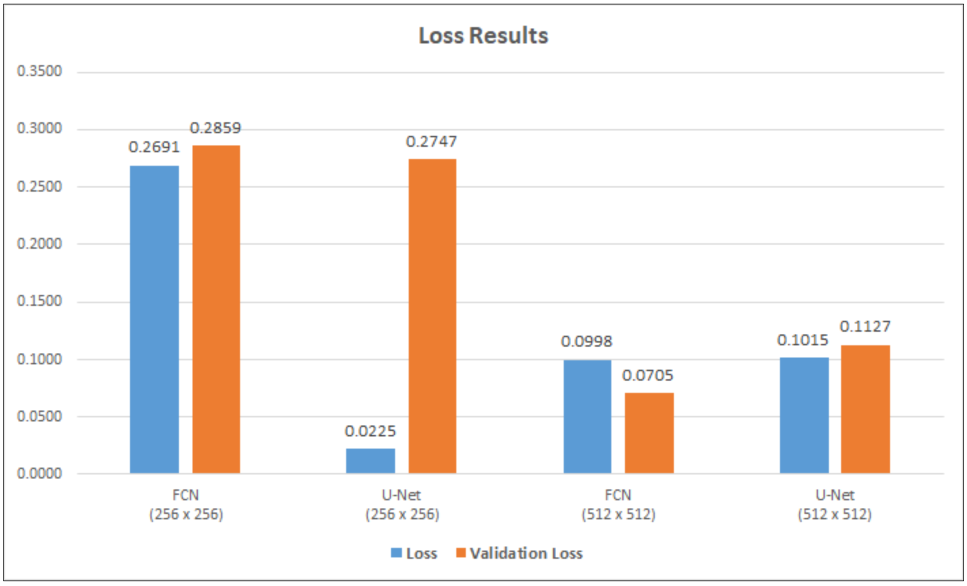


Figure 11. Training and validation loss results of the models

Conclusions

Most of the recent studies especially realized in developed countries, regular areas such as urban have been selected as the test area. Using these data set as the training data may cause several significant problems in developing countries such as Turkey. This may result within severe problems of extracting the linear features such as roads both in rural and urban areas. For this reason, the data sets and rules which are selected to be used for training data set should be carefully examined and created if necessary for every country separately according to their land use/cover characteristics.

It was observed that U-Net architecture gave significantly more accurate results than FCN when using a 256 x 256 dataset. At the same time, FCN gave better but not significantly higher results than U-Net when trained with a 512 x 512 dataset. It can also be seen from the predicted segmentation results.

Acknowledgments

The authors would like to thank Volodymyr Mnih (University of Toronto) for his permission to use his labelled data set.

References

Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481-2495.

Barghout, L., Lee, L. (2004). U.S. Patent Application No. 10/618,543.

Çelik, İ., Gazioğlu, C. (2020). Coastline Difference Measurement (CDM) Method, *International Journal of Environment and Geoinformatics,* 7(1): 1-5. doi. 10.30897/ijegeo.706792.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv*:1412.7062.

Cheng, D., Liao, R., Fidler, S., Urtasun, R. (2019). Darnet: Deep active ray network for building segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7431-7439).

Cheng, G., Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117, 11-28.

Chollet, F. (2018). *Deep Learning with Python*. Manning.

Digitalogy (2020). The Difference Between Artificial Intelligence, Machine Learning, And Deep Learning. Retrieved 15 May 2020 from https://blog.digitalogy.co/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning/

Erdem, F., Avdan, U. (2020). Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery. International Journal of Environment and Geoinformatics (IJEGEO), 7(3): 221-227. DOI: 10.30897/ijegeo.

Esetlili, M., Bektas Balcik, F., Balik Sanli, F., Kalkan, K., Ustuner, M., Goksel, Ç., Gazioğlu, C., Kurucu, Y. (2018). Comparison of Object and Pixel-Based Classifications for Mapping Crops Using Rapideye Imagery: A Case Study of Menemen Plain, Turkey. *International Journal of Environment and Geoinformatics,* 5(2), 231-243. doi: 10.30897/ijegeo.442002.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187, 27-48.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv*:1704.04861.

Liu, W., Rabinovich, A., Berg, A. C. (2015). Parsenet: Looking wider to see better. *arXiv preprint arXiv:*1506.04579

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D. (2020). Image Segmentation Using Deep Learning: A Survey. *arXiv preprint arXiv*:2001.05566.

Mnih, V. (2013). *Machine learning for aerial image labelling* (PhD thesis). University of Toronto, Toronto, Canada.

Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: *State of the art and research challenges. Expert Systems with Applications*, 105, 233-261.

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *In International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241). Springer, Cham.

Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234-241.

Wang, G., Li, W., Ourselin, S., Vercauteren, T. (2017, September). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In International MICCAI brainlesion workshop (pp. 178-190). Springer, Cham.