FEATURE EXTRACTION FROM SATELLITE IMAGES USING SEGNET AND FULLY CONVOLUTIONAL NETWORKS (FCN)

B. Sarıtürk[a] *, B. Bayram[b], Z. Duran[a], D. Z. Şeker[a]
[a] Istanbul Technical University, Faculty of Civil Engineering, Department of Geomatics Engineering, 34469 Maslak Istanbul, Turkey – (sariturkb, seker, duranza)@itu.edu.tr
[b] Yildiz Technical University, Faculty of Civil Engineering, Department of Geomatics Engineering, 34220 Davutpasa Istanbul, Turkey – (bayram@yildiz.edu.tr)

Key Words: Photogrammetry, Deep Learning, Feature Extraction, SegNet, Fully Convolutional Networks

ABSTRACT:

Object detection and classification are among the most popular topics in Photogrammetry and Remote Sensing studies. With technological developments, a large number of high-resolution satellite images have been obtained and it has become possible to distinguish many different objects. Despite all these developments, the need for human intervention in object detection and classification is seen as one of the major problems. Machine learning has been used as a priority option to this day to reduce this need. Although success has been achieved with this method, human intervention is still needed. In traditional machine learning techniques, it is necessary to extract feature vectors by experts to define the model and establish the system and these processes take a long time. Deep learning provides a great convenience by eliminating this problem. Deep learning methods carry out the learning process on raw data unlike traditional machine learning methods. Although deep learning has a long history, the main reasons for its increased popularity in recent years are; the availability of sufficient data for the training process and the availability of hardware to process the data. In this study, a performance comparison was made between two different artificial neural network architectures (SegNet and Fully Convolutional Networks (FCN)) which are used for object segmentation on high-resolution satellite images. These two different systems were trained using the same training dataset and their performances have been evaluated using the same test dataset. The creation of models and object extraction processes were performed on Python environment. The results show that there is not much significant difference between these two architectures in terms of accuracy, but FCN architecture is more successful than SegNet. However, this situation may vary according to the dataset used during the training of the system. Studies are underway to increase the performance of architectures.

Anahtar Sözcükler: Fotogrametri, Derin Öğrenme, Obje Çıkarımı, SegNet, Tam Konvolüsyonel Ağlar

ÖZET:

Fotogrametri ve Uzaktan Algılama çalışmalarında obje tespiti ve sınıflandırma, üzerinde çalışılan en güncel konulardandır. Teknolojinin gelişmesi ile birlikte, yüksek çözünürlüklü çok sayıda görüntü elde edilmeye başlanmış ve birçok farklı objenin ayırt edilebilmesi mümkün hale gelmiştir. Tüm bu gelişmelere rağmen görüntülerden obje tespiti ve sınıflandırma konusunda insan müdahalesine duyulan gereksinim önemli sorunlardan biri olarak görülmektedir. Bu ihtiyacın azaltılması için makine öğrenmesi, bugüne kadar öncelikli bir seçenek olarak kullanılmıştır. Bu yöntemle başarılar elde edilmiş olsa da, insana bağlılık halen devam etmektedir. Geleneksel makine öğrenmesi tekniklerinde, model tanımlamak ve sistemi kurmak için uzman kişilerce öznitelik vektörlerinin çıkarılması gerekmekte ve bu işlemler uzun zaman almaktadır. Derin öğrenme bu sorunu ortadan kaldırarak büyük bir kolaylık sağlamaktadır. Derin öğrenme yöntemleri, geleneksel makine öğrenmesi yöntemlerinin aksine öğrenme işlemini ham veri üzerinde yapmaktadırlar. Derin öğrenmenin uzun bir geçmişi olsa da, popülerliğinin son yıllarda artmasının temel sebepleri; eğitim işlemi için yeterli verinin kolaylıkla elde edilebiliyor olması ve bu veriyi işleyecek donanımların mevcut olmasıdır. Bu çalışmada, yüksek çözünürlüklü uydu görüntüleri üzerinden obje segmentasyonu için kullanılan iki farklı yapay sinir ağı mimarisi (SegNet ve Tam Konvolüsyonel Ağlar - Fully Convolutional Networks (FCN)) kullanılarak bu iki model arasında bir performans karşılaştırması gerçekleştirilmiştir. Bu iki farklı sistem, aynı eğitim veri seti kullanılarak eğitilmiş ve aynı test veri seti kullanılarak performansları test edilmiştir. Modellerin oluşturulması ve obje çıkarımı işlemleri Python ortamında gerçekleştirilmiştir. Sonuçlara bakıldığında, iki mimari arasında doğruluk açısından çok fazla fark bulunmadığı fakat FCN mimarisinin SegNet'e göre daha başarılı olduğu görülmüştür. Ancak bu durum, sistemin eğitimi sırasında kullanılan veri setine göre farklılıklar gösterebilmektedir. Mimarilerin performanslarının arttırılması için çalışmalara devam edilmektedir.

1. INTRODUCTION

Building detection from satellite remote sensing and photogrammetric data has been one of the most challenging tasks with important research and development efforts during the last decades (Vakalopoulou et al., 2015). In the field of remote sensing, for applications such as urban planning, land use analysis and automatic updating or generation of maps, automatic extraction of building outlines is a long-standing problem.

Buildings, which serve as the most significant place for human livelihood, are key elements on digital mapping of urban areas. With the rapid urban development, tremendous efforts are continually allocated to creating and updating location information of buildings for various fields. Aerial photogrammetry has been an effective technology for accurate mapping of buildings due to its capability for high-resolution imaging over large-scale areas. Unfortunately, automatic mapping of buildings is still limited by the insufficient

detection/segmentation accuracy on aerial images. Most cases require considerable amounts of manual intervention (Chen et al., 2018).

Recent years, based on the rapid development of imaging sensors and operating platforms, a dramatic increase in the availability and accessibility of very high resolution (VHR) remote sensing imagery has made this problem increasingly urgent [1]. Extracting buildings directly from images containing various backgrounds is very challenging because of the complexity of color, luminance and texture conditions.

Recent progress in computer vision (CV) field indicates that, with support from sufficient computing power and large training datasets, deep learning techniques such as Convolutional Neural Network (CNN) (LeCun et al., 1989) can substantially improve the performance of object detection and semantic segmentation from first-person or ground-level imagery (He et al., 2016; Krizhevsky et al., 2012). This condition strongly suggests that deep learning will play a critical role in promoting the accuracy of building detection toward practical applications of automatic mapping.

Since AlexNet overwhelmingly won the Large Scale Visual Recognition Challenge 2010 (LSVRC-2010) and 2012 (Krizhevsky et al., 2012) and based on the availability of open-source large-scale annotated datasets, CNN-based algorithms have become the gold standard in many computer vision tasks, such as image classification, object detection, and image segmentation. Initially, researchers mainly applied patch-based CNN methods to detecting or segmenting buildings in aerial or satellite images and significantly improved classification performance. However, owing to extreme memory costs and low computational efficiency, Fully Convolutional Networks (FCNs) have recently attracted more attention in this area (Wu et al., 2018).

In this study, a comparison was made between SegNet and Fully Convolutional Networks (FCN) architectures. Inria Aerial Image Labeling Dataset which consists of 180 training images (with corresponding labels) and 180 test images was used. These two different systems were trained using this dataset and their performances have been. The creation of models and object extraction processes were performed on Python environment on Google Colab.

2. DATASET AND METHODOLOGY

2.1 Dataset

Dataset selected to be used is “Inria Aerial Image Labeling Dataset”. The Inria Aerial Image Labeling addresses a core topic in remote sensing: the automatic pixel-wise labeling of aerial imagery.

Dataset features:

- Coverage of 810 km² (405 km² for training and 405 km² for testing),
- Aerial orthorectified color imagery with a spatial resolution of 0.3 m,
- Ground truth data for two semantic classes: building and not building (publicly disclosed only for the training subset)

The images cover dissimilar urban settlements, ranging from densely populated areas (e.g., San Francisco’s financial district) to alpine towns (e.g., Lienz in Austrian Tyrol) (Figure 1).

Figure 1. Dataset images and corresponding label images

Instead of splitting adjacent portions of the same images into the training and test subsets, different cities are included in each of the subsets. For example, images over Chicago are included in the training set (and not on the test set) and images over San Francisco are included on the test set (and not on the training set). The ultimate goal of this dataset is to assess the generalization power of the techniques: while Chicago imagery may be used for training, the system should label aerial images over other regions, with varying illumination conditions, urban landscape and time of the year.

This dataset is available for Inria Aerial Image Labeling Contest. The training set contains 180 color image tiles of size 5000×5000, covering a surface of 1500 m×1500 m each (at a 30 cm resolution). There are 36 tiles for each of the following regions:

- Austin (TX, USA)
- Chicago (IL, USA)
- Kitsap County (WA, USA)
- Western Tyrol (Austria)
- Vienna (Austria)

The format is GeoTIFF. The label data is in a different folder and the file names correspond exactly to those of the color images. In the case of the label data, the tiles are single-channel images with values 255 for the building class and 0 for the not building class.

The test set contains the same amount of tiles as the training set (but the label data is not disclosed). There are 36 tiles for each of the following regions:

- Bellingham (WA, USA)
- Bloomington (IN, USA)
- Innsbruck (Austria)
- San Francisco (CA, USA)
- Eastern Tyrol (Austria)Architectures

2.2 Methodology

2.2.1 SegNet: Segnet architecture is a CNN architecture designed to make deep learning algorithms more suitable for image segmentation. This architecture is illustrated in figure 2 Architecturally, SegNet has an encoder network and a decoder network that works according to this encoder. In addition, it has a pixel-based classification layer. Encoder network comprises 13 convolution layers, corresponding to the first 13 convolution layers of the VGG16 network, which is a pre-trained network designed for object classification. Therefore, the training process can be started from the weights that have been trained for classification on large datasets. At the deepest encoder output, fully connected layers are eliminated to protect higher resolution feature maps. This significantly reduces the number of parameters in the SegNet encoder network compared to other architectures.

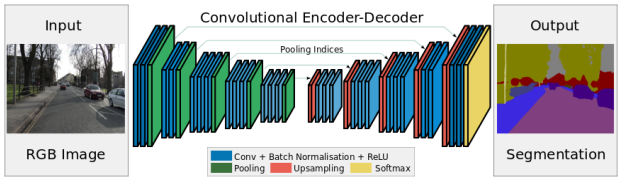

Figure 2. SegNet architecture (Badrinarayanan et al., 2017)

The novelty of SegNet lies is in the manner in which the decoder upsamples its lower resolution input feature map(s). Specifically, the decoder uses pooling indices computed in the maxpooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to upsample. The upsampled maps are sparse and are then convolved with trainable filters to produce dense feature maps. Badrinarayanan et al. (2017) compare proposed architecture with the widely adopted FCN and also with the well-known DeepLab-LargeFOV, DeconvNet architectures. This comparison reveals the memory versus accuracy trade-off involved in achieving good segmentation performance.

2.2.2 Fully Convolution Networks (FCN): Fully Convolutional Networks (FCNs) are being used for semantic segmentation of natural images, for multi-modal medical image analysis and multispectral satellite image segmentation (figure 3). Long et al. (2015) adapt contemporary classification networks (AlexNet, the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task.

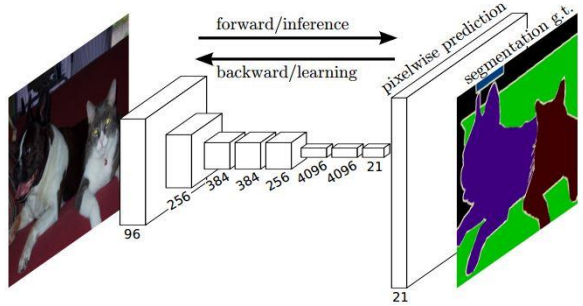

Figure 3. FCN end-to-end dense prediction pipeline (Long et al., 2015)

Then define a novel architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations (figure 4). Their fully convolutional network achieves state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes one third of a second for a typical image.

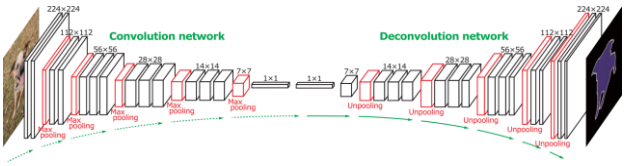

Figure 4. Encoding and decoding process of FCN (Noh et al., 2015)

3. STUDY

3.1 Pre-Processing

The dataset already divided into training and test regions. Later, the aerial imagery from training region, both RGB images and label images, are loaded as arrays, converted to gray scale (for ease of computations) and resized as 224 × 224 pixels to feed into the model. After that, data normalized by divided by 255.

3.2 Training

All training process was conducted on Google Colab. Google Colab is a free Jupyter notebook environment that allows users to use free Tesla K80 GPU. It runs in the cloud and stores its notebooks and data on Google Drive.

After pre-processing, training dataset split according to an 85% / 15% training / test ratio, ie. 153 and 27 respectively. The test dataset couldn't use in this step, because both RGB images and corresponding label images are necessary to train the network and test dataset only has RGB images.

Thereafter, architectures that used to train model (FCN and SegNet) defined in the system and training has been accomplished using these structures separately.

4. RESULTS

System that trained using FCN architecture has a %90.88 validation accuracy over 50 epochs (figure 5).

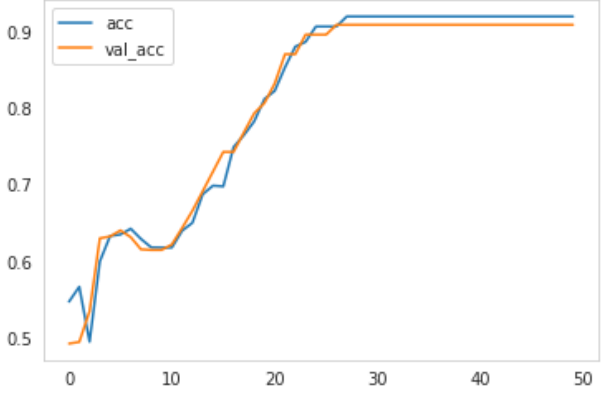

Figure 5. Accuracy and validation accuracy of FCN system

In addition, FCN trained system has 28.43% validation loss (figure 6).

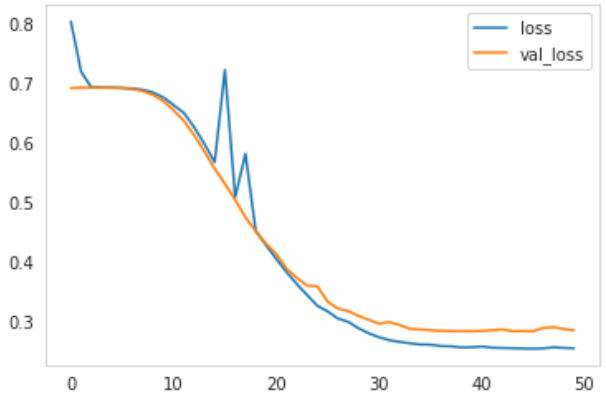

Figure 6. Loss and validation loss of FCN system

In this study, building detection was performed on Inria Aerial Labeling Dataset data set using deep learning architectures SegNet and FCN. 180 RGB images and corresponding label images from training set were used to train models.

The results show that there is not much significant difference between these two architectures in terms of accuracy, but FCN architecture is more successful than SegNet. However, this situation may vary according to the dataset used during the training of the system. Studies are underway to increase the performance of architectures.

The dataset used in this study can be considered as insufficient for a deep learning application. In order to overcome this and increase accuracy and efficiency of the models, dataset can be expended using data augmentation methods. However, this will also increase the load and processing time. With addition to that, pre-processing of data could be done more precisely to achieve predictions with higher accuracy. This study will be continued with trying different architectures and trying different hyperparameters.

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.

Budak, Ü. 2019. SegNet Mimarisi ile Bilgisayarlı Tomografi Görüntülerinden Karaciğer Bölgesinin Bölütlenmesi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 31(1), 215-222.

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., & Waslander, S. L. 2018. Aerial Imagery for Roof Segmentation: A Large-Scale Dataset Towards Automatic Mapping of Buildings. ISPRS Journal of Photogrammetry and Remote Sensing, 147, 42-55.

Chhor G., Aramburu C. B. & Bougdal-Lambert I. 2017. Satellite Image Segmentation for Building Detection Using U-Net. Stanford University.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems (pp. 1097-1105).

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation. doi:10.1162/neco.1989.1.4.541.

Lim, L. A., & Keles, H. Y. 2018. Learning Multi-Scale Features for Foreground Segmentation. arXiv preprint arXiv:1808.01477.

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).

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. 2017. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. IEEE International Symposium on Geoscience and Remote Sensing (IGARSS).

Noh, H., Hong, S., & Han, B. 2015. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1520-1528).

Vakalopoulou, M., Karantzalos, K., Komodakis, N., & Paragios, N. 2015. Building Detection in Very High Resolution Multispectral Data with Deep Learning Features. In 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 1873-1876). IEEE.

Wu, G., Guo, Z., Shi, X., Chen, Q., Xu, Y., Shibasaki, R., & Shao, X. 2018. A Boundary Regulated Network for Accurate Roof Segmentation and Outline Extraction. Remote Sensing, 10(8), 1195.