Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

DLRV Project Report

Semantic Segmentation of aerial images of forest scenery

Malika Navaratna, Urvashi Negi, Zain Ul Haq, Simon Deussen

Supervised by

Prof. Sebastian Houben

February 2022

1 Introduction

Loss of forest area in Germany is taking place at a high rate and due to factors like lack of rainfall, the spruce bark beetle is infesting and destroying forests in Germany. To overcome this and to achieve reforestation at a large scale, measures must be taken that can be completed quickly and automatically which would be faster than doing it manually.

Nowadays, unmanned Aerial Vehicles (UAVs) or drones are common in various applications such as aerial photography, agriculture, surveillance, product deliveries. UAVs can capture images by which users can generate high resolution images.

The above problem of loss of forest cover at a large scale has been identified and a project is in progress by a team at our university. The project is named Garrulus and details can be found on their website [1]. Garrulus is a project with the aim to reforest damaged forest area using UAVs. The prototype of the UAV would be surveying the terrain and would identify areas that are suitable for planting. Deep Learning is a field used for terrain surveillance and can provide relevant information. Image segmentation is a subpart of deep learning and is used to label the pixels of an image into different classes. Using image segmentation on aerial images captured by the drone is an approach to understand the terrain and plan the suitable land for reforestation.

1.1 Semantic Segmentation

Semantic segmentation refers to labelling each pixel of an image to a particular class. The figure 1 shows an example of a 2D image being classified into classes. In order to distinguish between forest and non forest area, semantic segmentation can be used to classify aerial images and decide the suitable area for reforestation. Architectures such as Fully Convolutional Neural Networks, U-Net are used for semantic segmentation.

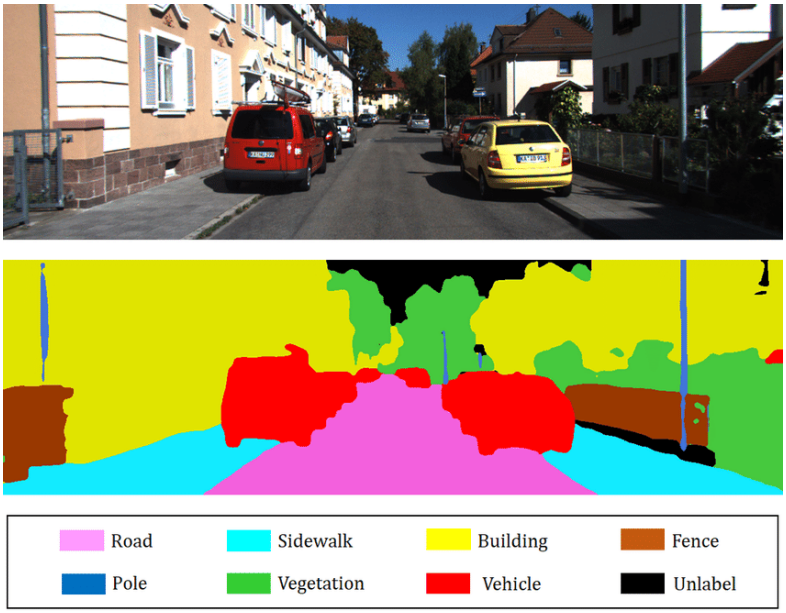

Figure 1: Semantic Segmentation as shown in [2]

2 Related Work

2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep learning architectures that use layers to implement convolutions on the images in order to extract relevant features. They consist of convolutional layer, pooling layer and a fully connected layer. The first breakthrough using CNNs was the AlexNet architecture [3] for the ImageNet challenge.

The architecture that further improved on this was the VGG architecture that made use of more deeper layers than AlexNet in order to get better results. [4]. The most well known architectures are the VGG-16 and VGG-19. The performance was further improved by Inception [5]. The inceptioon architecture went deeper and each layer had different convolutions to extract different features which were passed on the next layer with the help of a filter. But going deeper was only successful till a certain point and the performance was saturated. But the performance was improved by ResNet [6] where the authors did not go deeper but instead used skip connections in order to retain the "identity" or the relevant features.

2.2 Fully Convolutional Network

CNNs were designed for recognition of images and assigning a label to the image. Using the convolutional neural network for semantic segmentation was a bottleneck at the fully connected layer because this layer mixes the information from the entire image while getting the output. Therefore the convolutional neural network was modified for the application of semantic segmentation and called Fully Convolutional Network (FCN).[7] In this architecture, there is a downsampling path that extracts the relevant features, and an upsampling path which helps in localisation. Instead of the fully connected layer, there is a set of 1x1 kernel convolutions. FCNs employ skip connections to retain the information that was lost in the downsampling path. The authors mention that this helps in using images of arbitrary size.

2.3 U-Net

An example of an FCN is an architecture called U-Net. The name is because of the shape of the architecture as can be seen in the figure 2

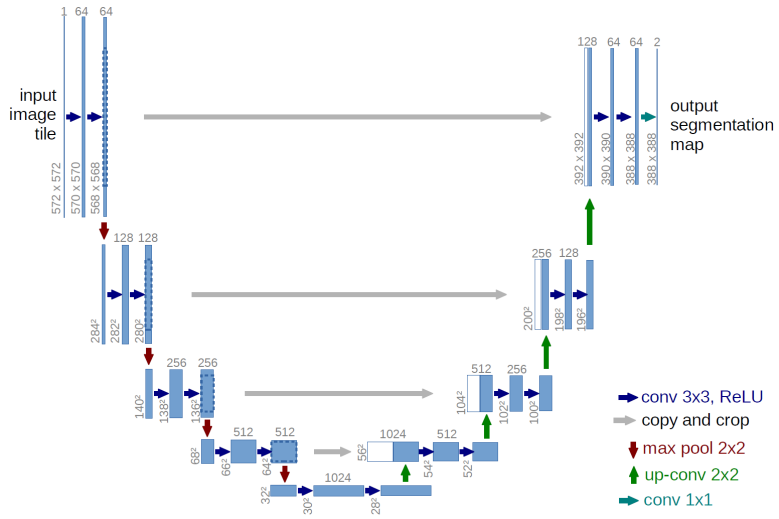

Figure 2: U-Net Architecture [8]

This architecture was initially designed for biomedical image segmentation and is an improvement on the FCN architecture. [8] It has become a popular choice for image segmentation as it requires fewer training images. This architecture

consists of an encoder and a decoder part and these are connected by a bridge in the bottom-most part of the image.

Encoder Network: This network extracts features with a sequence of encoder blocks. In the figure above this consists of a 3x3 convolution, then a ReLU activation function and then a max pooling layer. While going down the encoder or the contracting path, the dimensions are halved as compared to the previous layer and the feature channels are doubled.

Skip Connection: From the activation function of each layer of the encoder network, the output generated is used and concatenated to the corresponding layer of the decoder network. These connections are useful as they retain features that are useful in obtaining better semantic maps.

Decoder Network: In the decoder network, the semantic segmentation mask is generated. A 2x2 upscale convolution is performed. The skip connection is concatenated to each layer of the decoder network. Then two 3x3 convolutions are used and a ReLU activation function follows the skip connection. In this network, the dimensions are doubled while the feature channels are reduced by half. The last decoder goes through a 1x1 convolution with sigmoid activation. This function gives the segmentation mask respresenting the classification.

Bridge: This connects the encoder and decoder part. It has two 3x3 convolutions, and each convolution is followed by a ReLU activation function.

References

[1] "Garrulus." `https://www.h-brs.de/de/garrulus`.

[2] J. Jeong, T. Yoon, and J. Park, "Towards a meaningful 3d map using a 3d lidar and a camera," *Sensors*, vol. 18, p. 2571, 08 2018.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, p. 84–90, may 2017.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," pp. 3431–3440, 06 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.