A review of deep learning methods for semantic segmentation of remote sensing imagery

Dejan Ristovski

Faculty of computer science and engineering

Skopje, North Macedonia

Abstract

Semantic segmentation is one of many important problems in the field of computer vision. The segmentation of remote sensing images has wide range of applications such as urban planning and surveillance. With the introduction of deep learning models in machine learning, models for semantic segmentation problems has seen great improvements in performance. Every year new models are being developed and old models are getting improved. This research focuses on the fundamental models in deep learning for semantic segmentations and briefly explains what state of the art models look like. This research will help beginners in both semantic segmentation and remote sensing get a grasp of the basics needed for this field in machine learning.

1. Introduction

Computer vision has many forms of classification problems. While image classification focuses on classifying an image into a category, semantic segmentation is classifying every pixel in the image. This makes semantic segmentation one of the more problematic tasks in computer vision because the spatial information of each pixel is needed. The image can also contain objects that belong to different categories. This forces researches to invent new models that use methods both for extracting image features and upsampling features into useful data [1].

With the rise of deep neural networks, the field of semantic segmentation has seen many breakthroughs and traditional models used for this problem were forgotten. Researchers started using convolutional neural networks and modifying them into fully convolutional networks. The attention mechanism was introduced into this field and has given new ways to give context to the models [5]. This mechanism ignores irrelevant information and focuses on the relevant parts of the image. Later the transformer architecture was introduced in maybe one of the most popular papers in machine learning [6] and has been integrated into many state of the art models for semantic segmentation in the remote sensing field.

This research gives introduction into the basic models for semantic segmentation with both the supervised and unsupervised approach, and reviews their implementation into the remote sensing field. The rest of this article is arranged as follows. Section 2 gives the motivation for research in this filed. Section 3 describes the basic models for semantic segmentation and their application in remote sensing imagery. Section 4 gives brief description into the datasets used for performance analysis in the referenced papers. Section 5 concludes this research paper.

2. Motivation

Semantic segmentation has a important part in gathering information about the earths surface through remote sensing imagery. As technology increases, datasets with high-resolution satellite imagery are more available and demand is growing for extracting information from them. The application of semantic segmentation into remote sensing imagery plays a key role into many problems such as urban planning, economic assessment, resource management and environmental protection.

Deep learning models are becoming more and more complex, thus increasing the knowledge needed to get started with learning the best models for semantic segmentation. The amount of new research papers on deep learning published every year can be intimidating and push people away from the field of computer vision. According to [7], the number of papers based on remote sensing semantic segmentation has increased drastically since 2015 (See Figure 1).

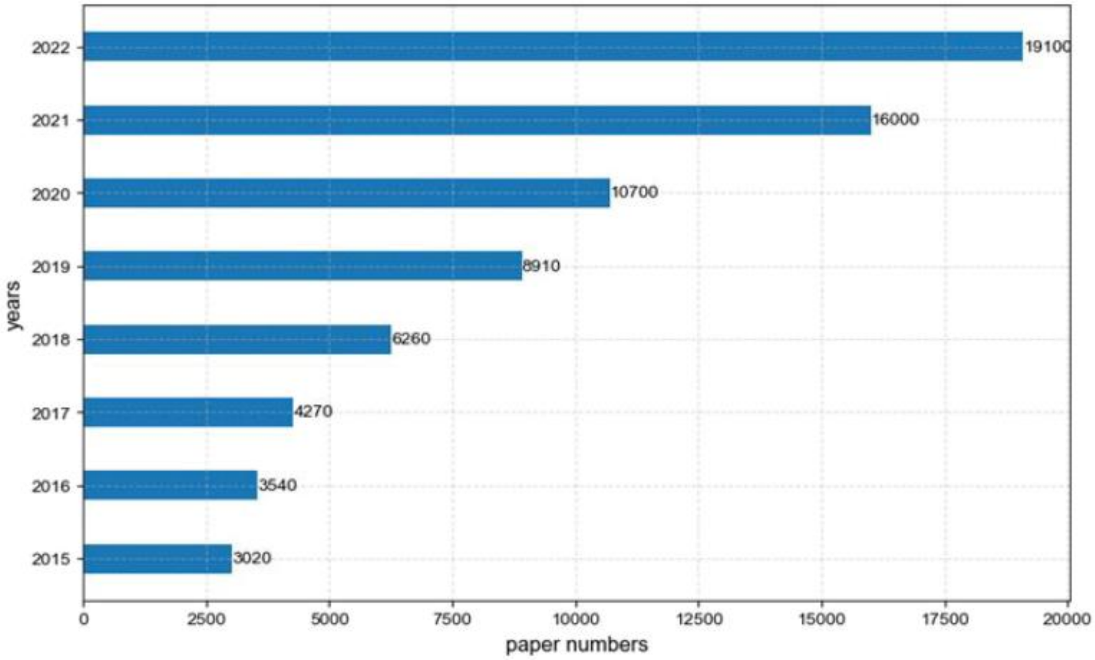

Figure 1. The number of papers based on remote sensing semantic segmentation since 2015. [7]

The motivation behind this research is from the need to help new researchers get started with the field of semantic segmentation. By explaining the basic models needed for semantic segmentation I intend to help beginners and also attract new researchers into this field of study. By the end of this research paper, the reader will have a strong foundation for deep learning semantic segmentation models and its applications into remote sensing imagery.

3. Semantic segmentation models

Many of the papers based on semantic segmentation are using CNN based models. This section describes some of the models including: fully convolutional networks (FCN), ensemble method and generative adversarial networks (GAN). A brief introduction into the transformer models with attention mechanisms for semantic segmentation is also included.

3.1. Fully Convolutional Network

Fully convolutional networks are very similar to basic CNNs. Basic CNNs are composed of convolutional layers, pooling layers and fully connected layers that flatten the features. Doing so, the model loses some vital information like the spatial coordinates of each pixel. This is useful in image classification where the model needs to predict only one class based on the entire image. However, when working with semantic segmentation, the spatial information for each pixel is essential. This problem has already been solved when the FCN was first introduced [1]. By replacing the fully connected layer of a CNN with a 1x1 convolution layer that covers the entire input region, a CNN model can be turned into a FCN. Doing so we add the spatial information needed for semantic segmentation. This process is called convolutionalization shown in Figure 2.

In CNNs the input dimension is fixed because of the fully connected layers. On the other hand, the FCN having only convolutional layers adds to the models flexibility. However, now the problem of upsampling the now downsampled image has arised. The output image needs to be of the same dimensions as the input image. This can be done by deconvolution also known as transposed convolution (See Figure 3). It is done by a reverse convolution and pooling operation which can recover the input size. This is not the ideal solution to the problem.

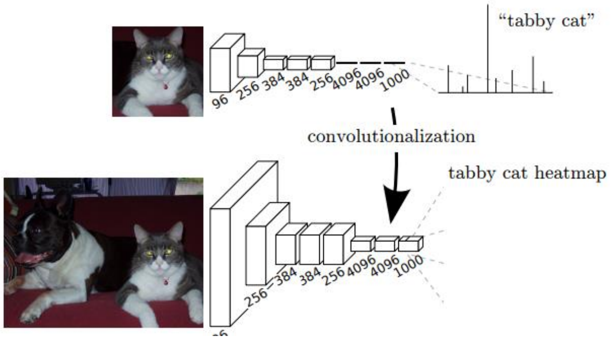

Figure 2. Convolutionalization [1]

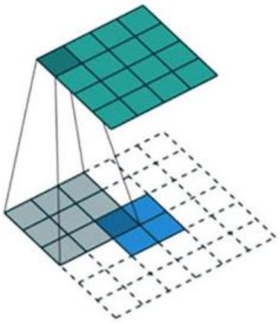

Figure 3. Deconvolution [7]

A better way to upsample a convoluted image was proposed in [1]. As shown in Figure 4 we can combine the upsampled coarse output with the previous outputs of the pooling layers. By doing this the model outputs a more precise image with the corresponding classes for each pixel. The more times we make this fuse operation, the better the segmented image will be. Based on the number of fuse operations [1] proposed 3 different models. The FCN-32s that is composed of only upsampling and no skip connections, the FCN-16s that is composed of a 2x upsampled output and the 4th pooling layer of the network and the FCN-8s that is composed of the 2x upsampled prediction from the FCN-16s combined with the 3rd pooling layer (see Figure 3).

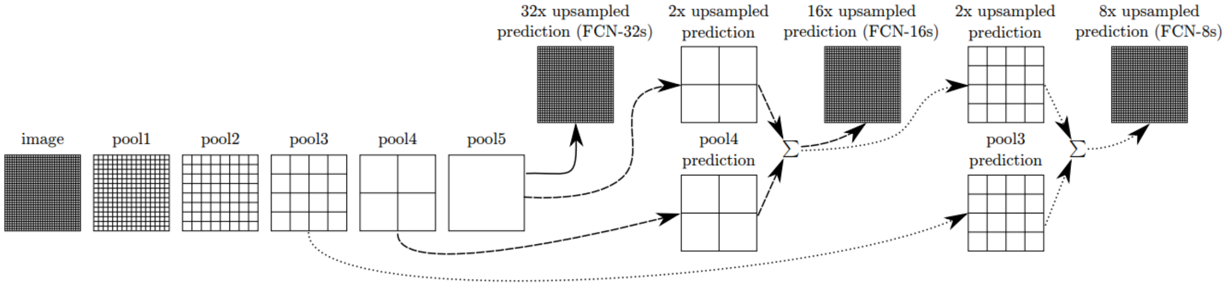

Figure 4. The process of transforming coarse outputs into an upsampled image [1]

The FCN architecture is the foundation of semantic segmentation and many models were implemented for remote sensing imagery. One such example is the ensemble model discussed in the next section.

3.2. Ensemble methods

FCNs need to be trained on large datasets to achieve satisfactory results. Luckly, semantic segmentation problems can be solved with fine tuning a pretrained model. Fine tuning is training an already trained model on new data in order to adjust the weights of the model to better fit the new data. Some of the more popular pretrained models using the FCN network are: FCN-ImageNet trained on ImageNet, FCN-Pascal trained on Pascal VOC and Places trained on the MIT Places database.

Ensemble methods are also present in semantic segmentation. The ensemble method combines the outputs of different models and fuses them into one output. This is useful in semantic segmentation because different models find different patterns in satellite imagery, thus their combination can result in better segmentation. In [2], the previously mentioned pretrained models are fine tuned on the ISPRS Vaihingen dataset (see Section 4 for more information) and combined as an ensemble model. The best performing model was obtained by combining the FCN-ImageNet and the FCN-Pascal model with an increase in performance over the basic FCN model.

3.3. Generative Adversarial Nets

The GAN architecture introduces a new way to find patterns in data. The architecture trains the model to generate new data by extracting information from datasets. This models works well with small datasets, thus being great for semantic segmentation problems.

GANs are composed of a generator and a discriminator. The generator generates new fake data as real as possible based on the training data and some random noise, and the discriminator tries to predict if the input is new or generated data. If the discriminator guessed right the parameters in the generator will be updated, and if the discriminator guessed wrong, then the discriminator will be updated. This model was first introduced in [3] and has since been used for semantic segmentation for a new way to find hidden patterns in remote sensing images and to work with small datasets.

In [4] a model based on only GANs was proposed. The discriminator in a standard GAN is a binary classifier that can be replaced with a fully convolutional multiclass classifier. Doing so makes the discriminator assign a class to each pixel instead of just trying to distinguish real or fake data. A softmax function is used to obtain the probability for each pixel belonging to a class. An additional class was also added for images that the discriminator classifies as fake. The architecture for the model proposed in [4] is shown in Figure 5.

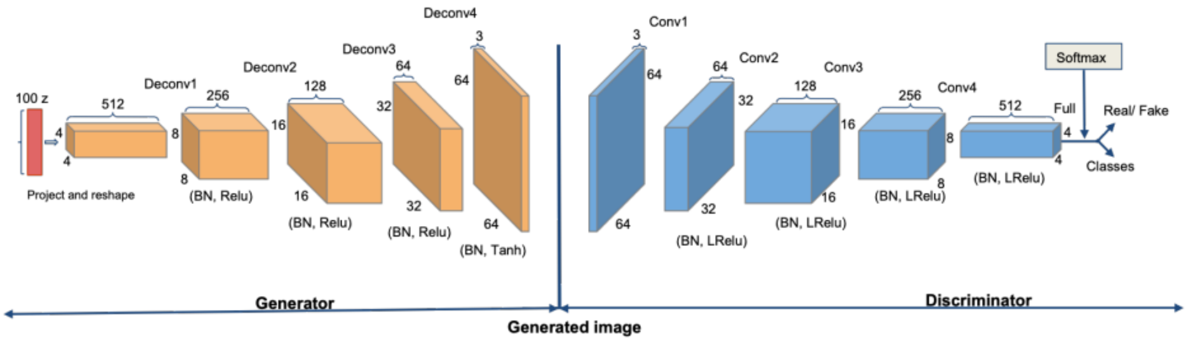

Figure 5. The network architecture of a semi-supervised GAN model [4]

The observed generated images in Figure 6 from the model trained on the ISPRS Vaihingen and the ISPRS Potsdam datasets clearly represent something unreal, but they also show that the model can find some hidden patterns in the data. Trees, buildings and cars are generated in the way the model understands them.

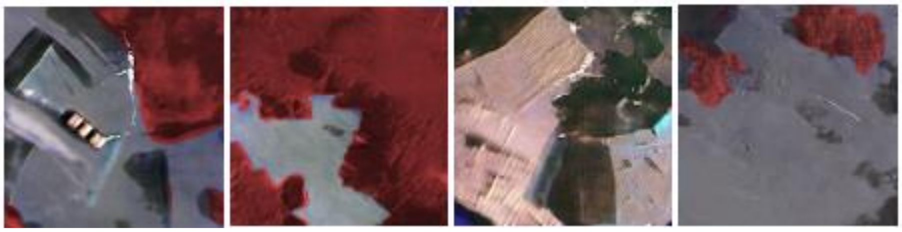

Figure 6. Generated images from the ISPRS Vaihingen and the ISPRS Potsdam datasets [4]

3.4. State of the art models

Since the appearance of the transformer model, models based on attention mechanism have been developed for semantic segmentation. The attention mechanism adds weights on different parts of the image, thus adding more context to the model and increasing the performance. This is also a very computational task meaning the model needs more time to train to learn all the weights for the image. However, their implementation is essential for achieving the best performance.

In [5] the attention mechanism was implemented in the skip connections for the upsampling of the image in U-Nets. The upsampling is done with the attention layer that takes the spatial information from the encoder layers and the feature information from the decoder. This is a small but significant change that improves the upsampling of the image, thus giving better results in semantic segmentation for remote sensing images.

According to [7] the best models for semantic segmentation are the models based on the attention mechanism and the transformer architecture. These models are a big improvement from the original FCNs with an increase in the mF1 score for the ISPRS Vaihingen and ISPRS Potsdam datasets.

4. Remote sensing datasets

4.1. ISPRS Vaihingen

The ISPRS Vaihingen dataset was captured over Vaihingen in Germany. Vaihingen is a small village with many independent buildings. The dataset consists of 33 images of consisting of a true orthophoto (TOP) with 2500 x 2000 pixels. 16 out of the 33 images are labeled with 6 classes: impervious ground, architecture, low vegetation, tree, car, and clutter.

4.2. ISPRS Potsdam

The ISPRS Potsdam dataset consists of 38 image tiles that cover 3.42 km^2 from the city Potsdam in Germany. The images are of 6000 x 6000 pixels and come in different channel compositions depending on the participants needs. Each channel has a spectral resolution of 8bits:

- IRRG: 3 channels (IR-R-G)
- RGB: 3 channels (R-G-B)
- RGBIR: 4 channels (R-G-B-IR)

This dataset is also labeled with the same 6 classes as the Vaihingen dataset.

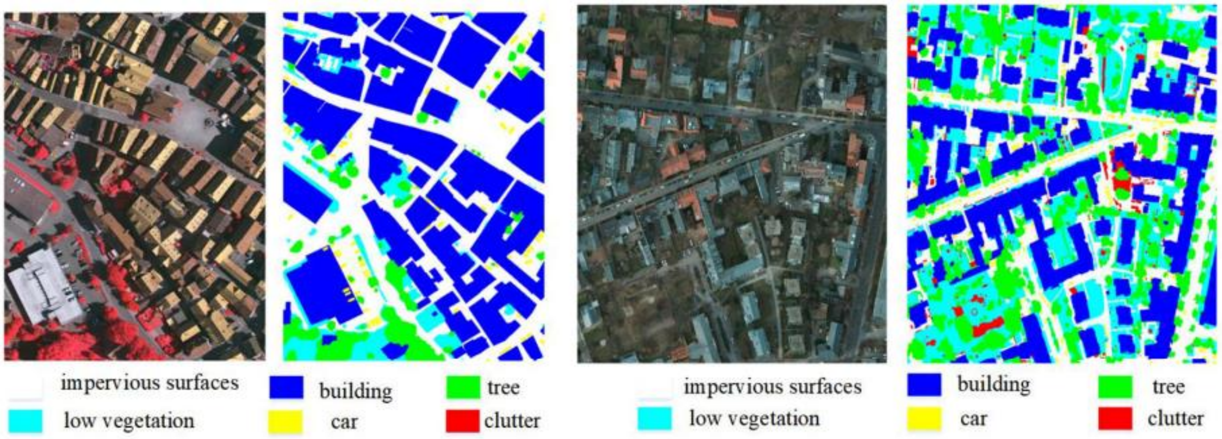

Figure 7. The ISPRS Vaihingen and ISPRS Potsdam datasets

5. Conclusion

This paper introduced the basic models for semantic segmentation. The FCN network was introduced that set a solid foundation. Then the GAN and ensemble models were implemented to fix the problem with small datasets and attention and transformer models were mentioned as state of the art. The transformer model gave big improvements in the semantic segmentation of remote sensing datasets such as the ISPRS Vaihingen and ISPRS Potsdam. However, there are still challenges in this field such as:

- The amount of papers released in this field can be intimidating making it harder for beginners to get started
- High resolution remote sensing images require manual pixel labeling, which is very labor-intensive
- Datasets have insufficient samples and require pretrained models to obtain satisfactory results
- Deep learning models for semantic segmentation typically require large amounts of computational resources and memory, especially for processing high-resolution remote sensing imagery

In the future researchers should focus on understanding the transformer architecture and its implementation in semantic segmentation to achieve the best possible results.

6. References

[1] Long, J., Shelhamer, E., and Darrell, T. (2019). "Fully convolutional networks for semantic segmentation" https://ieeexplore.ieee.org/document/7478072

[2] Marmanis, M., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U. (2016) "Semantic segmentation of aerial images with an ensemble of CNSS" https://www.researchgate.net/publication/312217868_Semantic_Segmentation_of_Aerial_Images_with_an_Ensemble_of_CNSS

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al.(2014), "Generative adversarial networks" https://arxiv.org/abs/1406.2661

[4] Kerdegari, H., Razaak, M., Argyriou, V., and Remagnino, P. (2019). "Urban scene segmentation using semi-supervised GAN" https://www.researchgate.net/publication/336343105_Urban_scene_segmentation_using_semi-supervised_GAN

[5] Ozan Oktay, Jo Schlemper, Loic Le Folgoc et al. (2018), "Attention U-Net: Learning Where to Look for the Pancreas" https://arxiv.org/abs/1804.03999

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. (2017), "Attention Is All You Need" https://arxiv.org/abs/1706.03762

[7] Jinna Lv , Qi Shen, Mingzheng Lv et al. (2023) "Deep learning-based semantic segmentation of remote sensing images" https://www.frontiersin.org/articles/10.3389/fevo.2023.1201125/full