Aerial Drone 2D Semantic Segmentation

Vibhanshu Jain

Prashant Jagwani

Abstract

The fundamental issue of semantic segmentation has lately gained attention in the disciplines of computer vision and machine learning. One of the crucial phases in developing complicated robotic systems, such as driverless cars/drones, human-friendly robots, robot-assisted surgery, and intelligent military systems, is to assign a separate class label to each pixel of a picture. So, it should come as no surprise that prominent businesses in the industry are now urgently tackling this issue alongside academic organizations that study artificial intelligence.

Models are trained for deep learning using a variety of applications and image datasets. Many studies are being conducted using convolutional operation-based methods in a variety of disciplines, including hand-arm identification in augmented reality, self-driving automobiles, drone-assisted aerial photography, and military technology. The human eye is capable of categorizing and separating what it perceives with ease. However, the difficulty of interpreting images, which is the equivalent of this skill in artificial intelligence technology, is covered under the heading of computer vision. The U-Net architecture, which was created for biological image segmentation and includes a real-world project that segments aerial imagery acquired by a drone using U-Net, is one of the segmentation techniques.

1. Problem Statement

Implementing a Encoder-Decoder architecture (U-Net) for a Semantic segmentation of Aerial Drone images for increasing the safety of autonomous drone flight and landing procedures. How to teach drone to see what is below and segment the object with high resolution is performed using Semantic Segmentation.

In a nutshell, semantic segmentation in computer vision is a pixel-wise labeling method. This can be performed using a U-Net architecture with transfer learning. We use a MobileNetV2 as the backbone architecture of the U-Net.

2. Proposed Solution

We developed a U-Net model that utilizes a MobileNetV2 backbone. Our goal was to improve the model's training and generalization performance, so we incorporated regularization techniques such as weight decay and image augmentation methods like Grid distortion and flipping.

Our modified U-Net architecture is designed to produce highly accurate segmentations using very few training images. We achieved this by adding layers to the typical contracting network, replacing pooling operators with upsampling operators to increase the output resolution. The high-resolution features from the contracting path are combined with the upsampled output to enable precise localization, and a convolution layer assembles the final output.

In our architecture, we increased the number of feature channels in the upsampling path to propagate context information to higher resolution layers, creating a symmetric and u-shaped network. We eliminated fully connected layers and only used the valid part of each convolution to generate a segmentation map containing pixels with full context from the input image.

We implemented an overlap-tile strategy to enable the network to seamlessly segment large images, extrapolating missing context by mirroring the input image. This tiling strategy allows us to apply the network to arbitrarily large images without limiting the resolution due to GPU memory constraints.

3. Dataset Details

Dataset Name: Semantic Drone Dataset

Description: The imagery depicts more than 20 houses from nadir (bird's eye) view acquired at an altitude of 5 to 30 meters above ground. A high resolution camera was used to acquire images at a size of 6000x4000px (24Mpx). The training set contains 400 publicly available images and the test set is made up of 200 private images.

We prepared pixel-accurate annotation for the same training and test set.

Table 1: Semantic classes of the Drone Dataset

- tree
- gras
- other vegetation
- dirt
- gravel
- rocks
- water
- paved area
- pool
- person
- dog
- car
- bicycle
- roof
- wall
- fence
- fence-pole
- window
- door
- obstacle

4. Survey

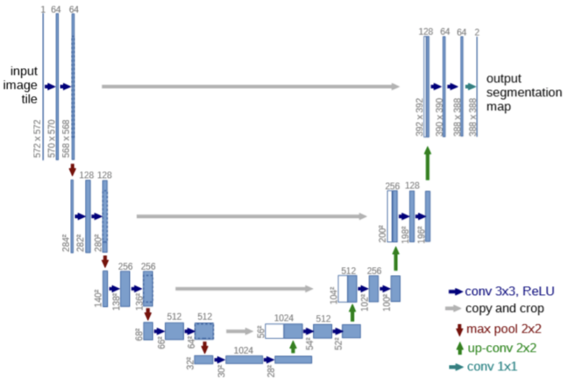
The paper titled ”A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D images” aims to provide a comprehensive overview of the evolution of semantic segmentation architectures and the chronological order of techniques in the context of new challenges. Semantic segmentation is a rapidly evolving field with significant interest from researchers and practitioners alike, and many survey studies have been published in recent years. However, the authors of this paper argue that many of these surveys lack an overarching vision or necessary depth of analysis regarding deep learning-based methods.

To address this gap, the authors provide a detailed analysis of the chronological evolution of semantic segmentation techniques based on deep learning. They present a comprehensive survey of related methods in a tabular format and explain them briefly in chronological order, providing their metric performance and computational efficiency. By doing so, they hope to provide readers with a clear understanding of the current state-of-the-art and future directions of 2D semantic segmentation.

This paper provides an important contribution to the field of semantic segmentation by presenting a detailed overview of the evolution of deep learning-based architectures for 2D image segmentation. It offers a comprehensive and structured survey of related techniques that will be valuable for researchers and practitioners in this field. The paper's detailed analysis and chronological order of techniques will help readers better understand the development of semantic segmentation and provide insights into future research directions.

5. Model Architecture

We utilized a U-Net architecture for our model, which comprises an encoder-decoder structure with skip connections that connect the same levels of encoder and decoder, culminating in an input-sized classification layer. This design results in an efficient computational load, as it does not involve fully connected layers or a refinement block. Additionally, we incorporated a MobileNetV2 as the backbone architecture of the U-Net, which handles the upsampling and downsampling operations internally.



The U-Net architecture is a popular and effective model for image segmentation, consisting of a contracting path on the left side and an expansive path on the right side. The contracting path follows the standard architecture of a convolutional network, with repeated application of 3x3 convolutions, ReLU activations, and 2x2 max pooling operations for downsampling. At each downsampling step, the number of feature channels is doubled to extract more abstract features.

In the expansive path, the feature map is upsampled followed by a 2x2 convolution that halves the number of feature channels. The upsampled feature map is then concatenated with the corresponding cropped feature map from the contracting path, followed by two 3x3 convolutions with ReLU activations. The cropping is necessary to prevent the loss of border pixels during convolutions. Finally, a 1x1 convolution is used to map each 64-component feature vector to the desired number of classes, resulting in a total of 23 convolutional layers in the network.

The MobileNetV2 architecture is used as the backbone of the U-Net, which performs the upsampling and downsampling by itself. The MobileNetV2 architecture is based on an inverted residual structure, where the input and output of the residual block are thin bottleneck layers. Unlike traditional residual models that use expanded representations in the input, MobileNetV2 uses lightweight depthwise convolutions to filter features in the intermediate expansion layer. This approach leads to an efficient and effective model for image segmentation.

In addition, it is important to remove non-linearities in the narrow layers of the network to maintain representational power. The authors demonstrate that this improves performance and provide an intuition that led to this design. Finally, the authors' approach allows for the decoupling of the input/output domains from the expressiveness of the transformation, providing a convenient framework for further analysis of the model's performance and behavior.

6. Training and Evaluation Techniques

We defined the following parameters for training:-

Training Parameters:

1. Loss: Cross-entropy loss, which computes the loss between the input and target for a classification problem with C classes. It is especially useful for an unbalanced training set.

2. Optimizer: AdamW, a variant of Adam optimizer that has an improved implementation of weight decay. Weight decay is a form of regularization that reduces the risk of overfitting.

3. Learning Rate: 1e-3

4. Weight Decay: 1e-4. Weight decay is a regularization technique that involves adding a small penalty, usually the L2 norm of the weights, to the loss function.

5. Scheduler: This sets the learning rate of each parameter group according to the 1cycle learning rate policy. The 1cycle policy anneals the learning rate from an initial value to a maximum learning rate and then back to a minimum value lower than the initial learning rate. Evaluation Parameters:

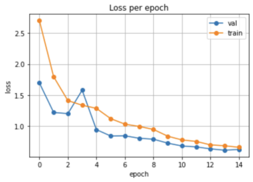
We utilized some of the best evaluation methods outlined in ”A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D Images” paper, which include:

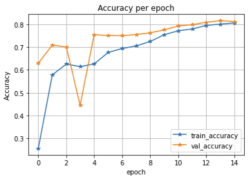
1. Pixel Accuracy: A simple metric that calculates the ratio of properly classified pixels to their total number.

2. mIou (mean Intersection over Union): This is the class-averaged IoU (Intersection over Union) also known as the Jaccard Index. IoU is a statistical measure used to compare the similarity and diversity of sample sets. In semantic segmentation, it is the ratio of the intersection of the pixel-wise classification results with the ground truth, to their union.

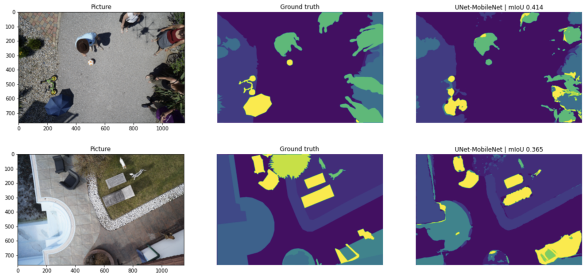
7. Results

After training the model, we get the below loss and accuracy:-





Predictions



8. Conclusion

The current model shows good performance on the test images, with the generated mask appearing similar to the ground truth. However, this solution is just a basic implementation and there is room for improvement using various methods.

To enhance the mean intersection over union (MIOU) score, we can explore different backbone architectures like efficientnets. Additionally, we can try using different architectures such as FPN (feature pyramid networks), or object detection methods like RCNN and DeeplabV3.

Further improvements can be achieved by performing hyper-parameter tuning and incorporating more image augmentations into the process.

9. Future Work

Aerial Drone 2D Semantic Segmentation is a rapidly developing field in computer vision, and there is still a lot of room for improvement in terms of accuracy and efficiency. One potential avenue for future work is exploring different architectures beyond Resnet, Efficientnet, Resnext, and Xceptionet. For example, there has been recent interest in using transformer-based models for semantic segmentation tasks, which have shown promising results in other computer vision tasks such as image classification and object detection. Additionally, there is potential for combining multiple architectures to create an ensemble model that leverages the strengths of each individual architecture.

Another area for future work is developing techniques for efficient inference on resource-constrained devices such as drones. While some of the aforementioned architectures have been optimized for efficiency, there is still a need for more lightweight models that can run in real-time on low-power devices. This could involve exploring techniques such as knowledge distillation or pruning to reduce model size without sacrificing accuracy, as well as developing hardware-specific optimizations for popular drone platforms. Ultimately, the goal of future research in this area should be to create models that can perform accurate semantic segmentation on aerial drone imagery while minimizing the computational and power requirements.

References

1. Olaf Ronneberger, Philipp Fischer, Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation (2015)

2. Irem Ulku, Erdem Akagunduz: A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D images (2019)

3. Dataset Name: Semantic Drone Dataset Dataset Link: http://dronedataset.icg.tugraz.at