Aerial Drone 2D Semantic Segmentation

Vibhanshu Jain Github Project 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	▪ rocks	▪ dog	▪ fence
▪ gras	▪ water	▪ car	▪ fence-pole
▪ other vegetation	▪ paved area	▪ bicycle	▪ window
▪ dirt	▪ pool	▪ roof	▪ door
▪ gravel	▪ person	▪ wall	▪ obstacle

4. Survey

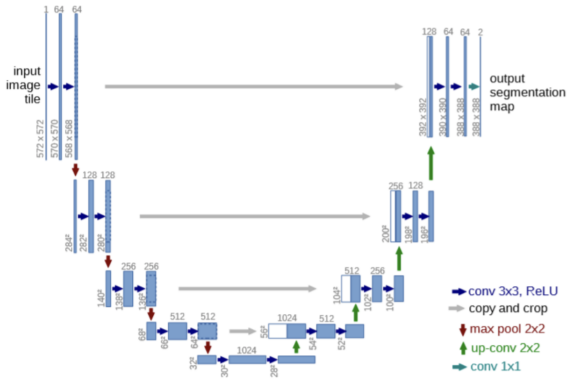
The paper titled "A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D images" aims to provide a comprehensive overview of the evolution of semantic segmentation architectures and the chronological order of techniques in the context of new challenges. Semantic segmentation is 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.

5.1. U-Net

U-Net is a popular neural network architecture for semantic segmentation, introduced in 2015 by Olaf Ronneberger, Philipp Fischer, and Thomas Brox. The U-Net architecture consists of an encoder-decoder structure, where the encoder is a series of convolutional layers that downsample the input image, and the decoder is a series of transposed convolutional layers that upsample the feature maps back to the original size. The novelty of U-Net lies in the skip connections that connect the corresponding layers in the encoder and decoder, which allows the network to preserve fine-grained details while also capturing high-level semantic information.

The U-Net architecture is particularly well-suited for biomedical image segmentation, where the task is to segment the different regions of interest in medical images such as MRI or CT scans. This is because the U-Net architecture can handle images with different resolutions and shapes, and it can also capture the fine details of the anatomical structures. Moreover, the skip connections in U-Net can help to overcome the class imbalance problem that often arises in medical image segmentation, where the foreground class (i.e., the region of interest) is much smaller than the background class.

The U-Net architecture has been extended and modified in various ways to improve its performance on different segmentation tasks. For example, some researchers have added attention mechanisms to the U-Net architecture to focus on the relevant regions of the input image, while others have introduced residual connections or dense connections to improve the information flow between the encoder and decoder. Another modification of U-Net is the use of multi-scale inputs, where the network is trained on images with different resolutions or scales to improve its generalization to unseen data.

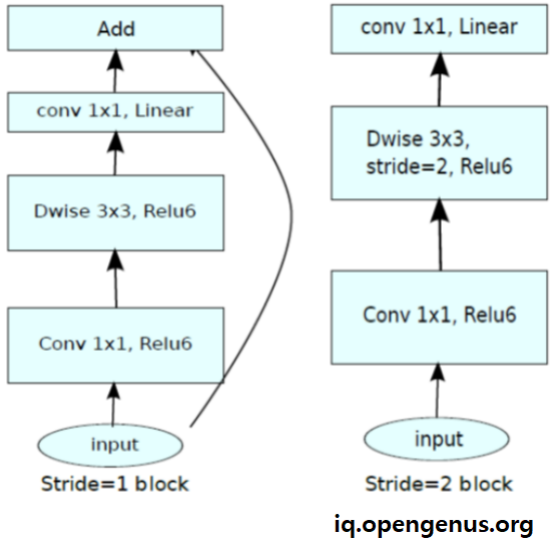
One of the key benefits of the U-Net architecture is its efficiency in terms of both computational complexity and memory usage. This is because U-Net does not use fully connected layers, which are computationally expensive, and it also uses convolutional layers with small kernel sizes, which are memory-efficient. Moreover, the skip connections in U-Net reduce the number of parameters needed to represent the input image, which further improves its efficiency.

Training U-Net for semantic segmentation typically involves minimizing a pixel-wise loss function, such as the mean squared error or cross-entropy loss, between the predicted segmentation map and the ground truth map. However, the class imbalance problem in biomedical image segmentation can lead to poor performance, especially for rare classes. To address this, some researchers have proposed using weighted loss functions that assign higher weights to the foreground class, or using dice loss, which is a measure of the overlap between the predicted and ground truth segmentation maps.

Overall, U-Net has become a popular and versatile architecture for semantic segmentation in various domains, including biomedical image segmentation, road segmentation, and building segmentation. Its efficiency, flexibility, and ability to preserve fine-grained details make it a powerful tool for image analysis and computer vision research.

5.2. MobileNetV2

MobileNetV2 is a neural network architecture that was introduced in 2018. It is a successor to the original MobileNet architecture, which was designed for mobile and embedded devices. MobileNetV2 improves upon the original by using a novel architecture that is based on an inverted residual structure. In this structure, the input and output of the residual block are thin bottleneck layers, which are followed by a linear bottleneck layer that expands the number of channels. 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 classification.



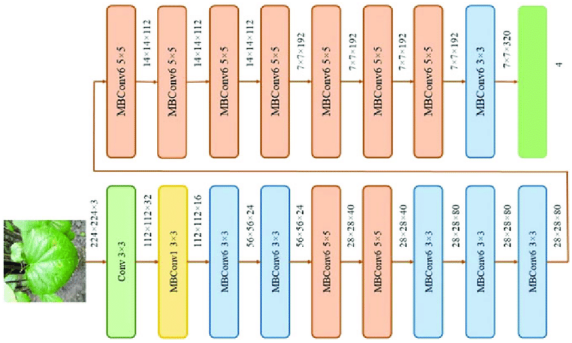
The depthwise separable convolution is a key component of MobileNetV2 that helps to reduce the number of parameters and computations in the model. In a standard convolutional layer, each filter operates on all channels of the input tensor. In contrast, a depthwise convolution applies a single filter to each channel of the input tensor. This reduces the number of parameters and computations by a factor equal to the number of input channels. MobileNetV2 also includes a pointwise convolution, which applies a 1x1 filter to combine the output of the depthwise convolution across channels. This step is used to increase the number of channels in the feature map, which enables the network to learn more

complex features.

MobileNetV2 is designed to be efficient and effective on mobile and embedded devices. The model is optimized for speed and memory usage, and it achieves state-of-the-art accuracy on benchmark datasets. The authors of the original paper demonstrate that MobileNetV2 outperforms other models such as Inception-V3 and ResNet-50 on ImageNet classification, while using fewer parameters and computations. MobileNetV2 has also been applied to other computer vision tasks such as object detection, semantic segmentation, and face recognition. Overall, MobileNetV2 is a powerful and versatile architecture that can be used in a wide range of applications where computational resources are limited.

5.3. Efficientnet

EfficientNet is a family of neural network architectures that was introduced in 2019. The key idea behind EfficientNet is to balance the trade-off between model size and accuracy by scaling the network's depth, width, and resolution in a principled way. The authors propose a novel scaling method that uses a compound scaling coefficient to increase the size of the network in a uniform way across all dimensions. This allows the model to achieve state-of-the-art accuracy on various computer vision tasks while using fewer parameters and computations than previous state-of-the-art models.

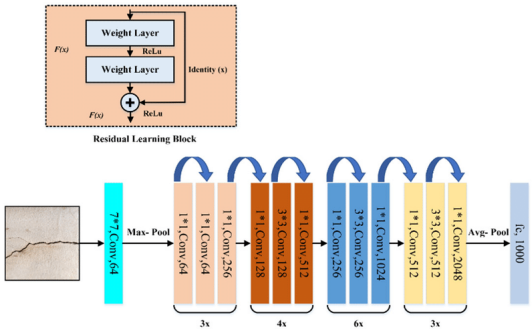


EfficientNet uses a backbone architecture that is based on a compound scaling of the popular MobileNetV2 architecture. The authors introduce a new type of block called the compound scaling block, which consists of a combination of depthwise separable convolution, squeeze-and-excitation, and a swish activation function. The depthwise separable convolution helps to reduce the number of parameters and computations, while the squeeze-and-excitation module helps to improve the model's expressive power. The swish activation function is a smooth and non-monotonic function that has been shown to outperform other activation functions such as ReLU.

EfficientNet has achieved state-of-the-art performance on various computer vision tasks, including image classification, object detection, and semantic segmentation. The authors demonstrate that EfficientNet achieves better accuracy than previous state-of-the-art models such as ResNet, Inception, and MobileNet, while using fewer parameters and computations. EfficientNet has also been shown to generalize well to new domains and datasets, indicating its potential as a general-purpose model for computer vision. Overall, EfficientNet is a powerful and versatile architecture that represents a significant advance in the field of neural network design

5.4. Resnet

Residual Networks (ResNets) are a type of deep neural network that was introduced in 2015 by Kaiming He et al. ResNets are built on the idea that very deep neural networks are difficult to train because of the vanishing gradients problem. The idea behind ResNets is to use shortcut connections, known as residual connections, to allow for gradients to flow more easily through the network, which can help to avoid the vanishing gradient problem.



One of the key features of ResNets is the use of residual blocks, which consist of a few layers that perform a nonlinear transformation on the input data, followed by a shortcut connection that skips over one or more layers. The residual connection allows the input to be added back to the output of the residual block, which can help to preserve information and improve gradient flow.

ResNets have been shown to be very effective for a wide range of computer vision tasks, such as image classification, object detection, and segmentation. In fact, ResNets have been the backbone of many state-of-the-art models in these areas. ResNets have also been applied to other domains, such as natural language processing and speech recognition, with promising results.

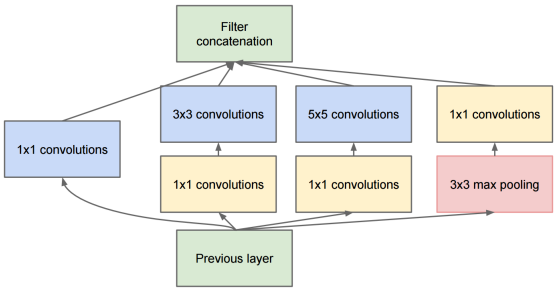
One of the advantages of ResNets is their ability to train very deep networks without sacrificing performance. By using residual connections, ResNets are able to mitigate the vanishing gradients problem, which can make it easier to train deeper networks. Additionally, ResNets are able to achieve state-of-the-art performance on many benchmark datasets, which has made them a popular choice for many computer vision tasks.

Overall, ResNets are an important development in the field of deep learning, and they have helped to push the state

of the art in computer vision and other domains. Their ability to train very deep networks and achieve state-of-the-art performance has made them a popular choice for many researchers and practitioners.

5.5. InceptionNet

InceptionNet, also known as GoogLeNet, is a deep neural network architecture that was introduced by Google researchers in 2014. The key idea behind InceptionNet is to use multiple parallel paths with different kernel sizes to extract features at different scales. This allows the network to capture both local and global features in an efficient way, leading to improved accuracy on various computer vision tasks. InceptionNet also introduced the concept of the "inception module", which consists of multiple parallel convolutional layers with different kernel sizes and pooling operations.



InceptionNet uses a deep and complex architecture with more than 22 layers, which makes it difficult to train without overfitting or vanishing gradients. To address this, the authors introduced several regularization techniques such as dropout and batch normalization. They also proposed a novel method called "label smoothing" to regularize the output of the network, which helps to prevent overconfidence in the predictions. In addition, InceptionNet uses a multi-level hierarchy of classifiers that allows the network to adapt to different levels of abstraction in the input data.

InceptionNet has achieved state-of-the-art performance on various computer vision tasks, including image classification, object detection, and semantic segmentation. The authors demonstrate that InceptionNet achieves better accuracy than previous state-of-the-art models such as AlexNet and VGG, while using fewer parameters and computations. InceptionNet has also been used as a building block in other neural network architectures such as Inception-ResNet and Inception-v4, which further improve the accuracy and efficiency of the network. Overall, InceptionNet represents a significant advance in the field of neural network design and has had a profound impact on the development of deep learning models for computer vision.

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

EfficientNet is based on a novel compound scaling method that uniformly scales all dimensions of depth, width, and resolution in a principled manner. This approach enables EfficientNet to achieve state-of-the-art accuracy on image classification tasks while using fewer parameters and computation than previous CNN architectures.

ResNet, on the other hand, introduced the concept of residual connections, which allows for better flow of gradients during training, leading to improved model accuracy.

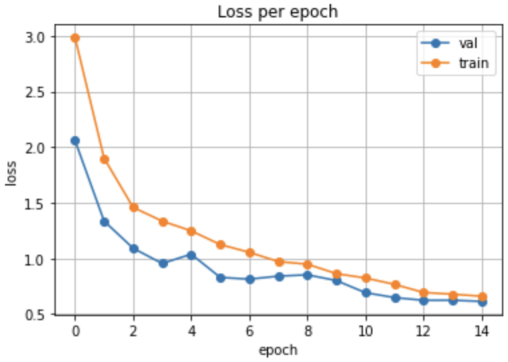
InceptionNet is another CNN architecture that uses multi-scale convolutional layers to capture different levels of features, which allows the model to better handle complex visual patterns.

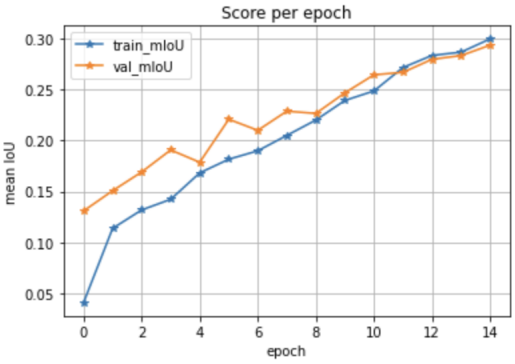
MobileNet was designed specifically for resource-constrained devices such as mobile phones and embedded systems. It uses depthwise separable convolutions, which are computationally efficient and reduce the number of parameters in the model.

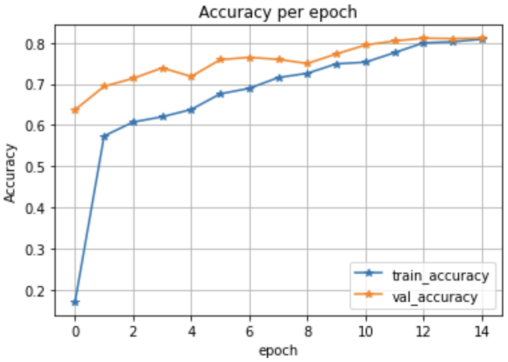
7.1. MobileNetV2

Graphs

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

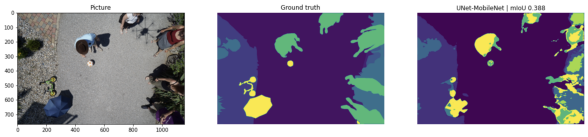






Predictions

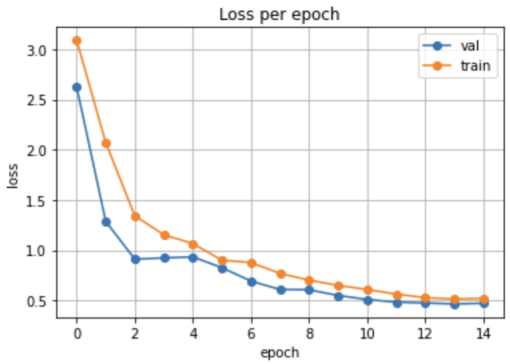
After training the model, we get the below prediction:-

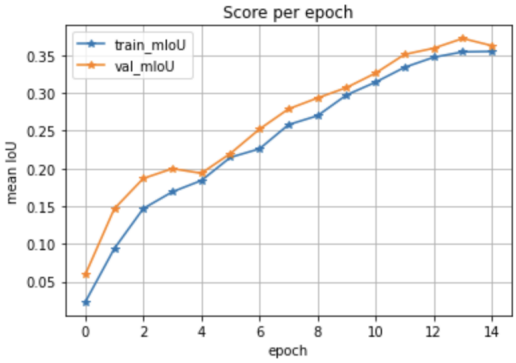


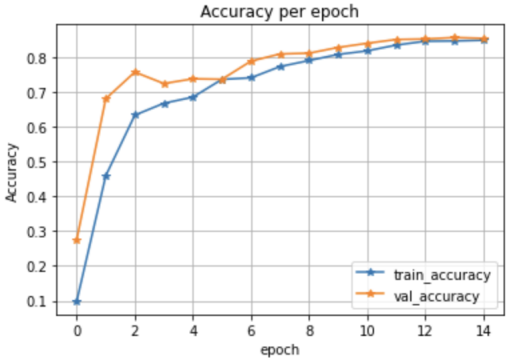
7.2. Efficientnet

Graphs

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

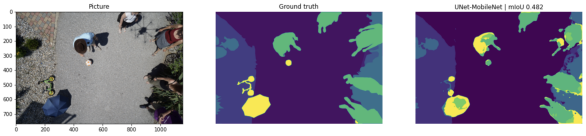






Predictions

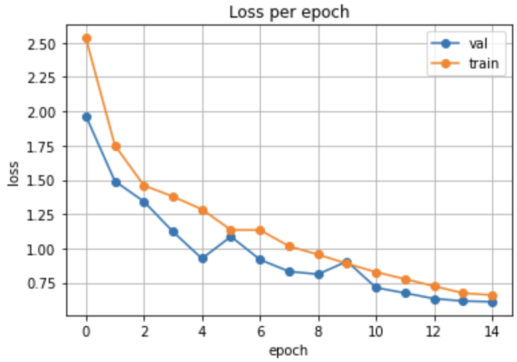
After training the model, we get the below prediction:-

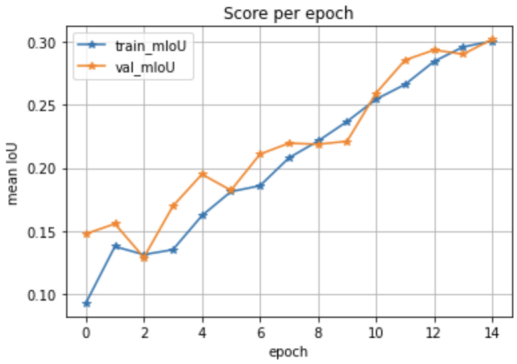


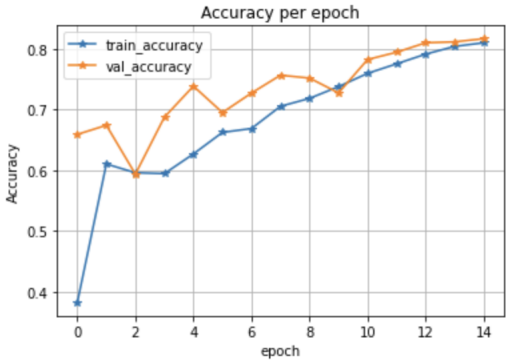
7.3. Resnet

Graphs

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

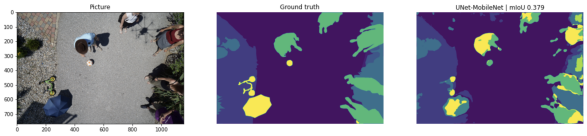






Predictions

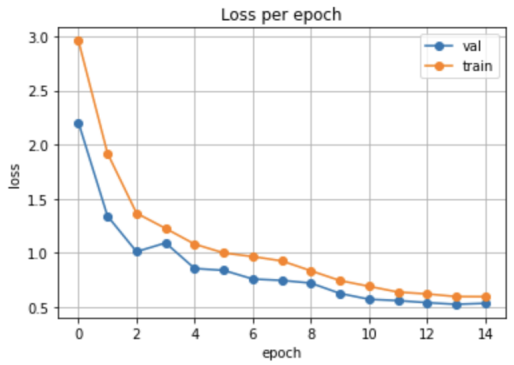
After training the model, we get the below prediction:-

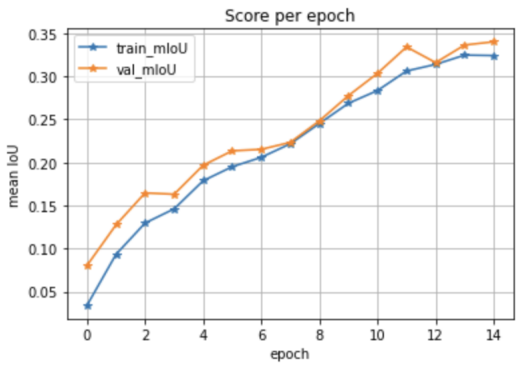


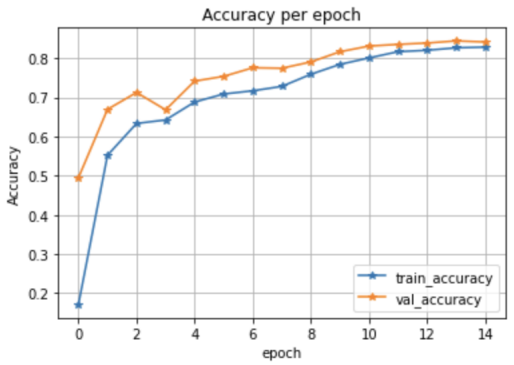
7.4. InceptionNetV2

Graphs

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

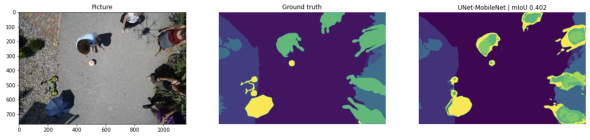






Predictions

After training the model, we get the below prediction:-



8. Conclusion

In conclusion, the choice of backbone architecture plays a critical role in the performance of image segmentation models. EfficientNet, with its compound scaling method, provides a powerful solution to efficiently scale up the model size while maintaining high accuracy, making it an attractive option for image segmentation tasks. ResNet's residual connections enable the flow of gradients during training, which helps improve the model's accuracy. InceptionNet's multi-scale convolutional layers are effective at handling complex visual patterns and have been successfully used in various image segmentation models.

MobileNet's design specifically caters to mobile and embedded devices by reducing the number of computations and parameters required, leading to faster inference times and lower latency. Furthermore, techniques like FPN, RCNN, and DeeplabV3 can significantly impact the performance of image segmentation models by improving the model's ability to capture complex spatial and semantic information.

In addition to exploring different architectures and techniques, hyper-parameter tuning and incorporating image augmentations can further improve the model's accuracy and robustness. Selecting the appropriate optimization technique based on the specific requirements of the task is critical to achieving optimal performance. Overall, image segmentation remains a challenging task, but by carefully selecting and implementing the right combination of techniques and architectures, it is possible to achieve state-of-the-art results

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