Land Cover Classification with CNN

Comparison of different CNN architectures for image segmentation

Claudia Ochoa, Eszter Pazmandi,
Sandra Vicaria

Term Paper for Analysis of Spatial Data and Images
Barcelona School of Economics
Submission Date: April 1, 2022

Abstract

In this project, we face the Land Cover Classification task of the DeepGlobe Satellite Image Understanding Challenge. DeepGlobe provides high-resolution satellite image datasets after signing up to the challenge. The land cover classification task of the DeepGlobe Challenge presents significant obstacles even to state of the art segmentation models due to a small amount of data, incomplete and sometimes incorrect labeling, and highly imbalanced classes. Our work consists of our intention to tackle this challenge with different image segmentation models. We demonstrate three approaches. First we apply the standard UNet architecture which is well-known to being used for semantic segmentation tasks. Second we use the VGG16 architecture as encoder on top of the UNet model and third we implement the DeepLabV3+ architecture, which is also known as a best practice in the field of segmentation tasks.

Keywords: semantic segmentation, neural network, UNet, autoencoder

1 Introduction

Satellite imagery is a powerful source of information as it contains more structured and uniform data compared to traditional images. It can provide very valuable information for agriculture, forestry, sustainable development, urban planning, prevention of emergencies and natural disasters, and much more. Government agencies and companies, such as DigitalGlobe, have access to very large satellite imagery nowadays, and as machine and deep learning methods dominate the computer vision field, public datasets and benchmarks have started to play an important role for relative scalability and reliability of different approaches.

As time goes Machine learning algorithms have improved substantially and image segmentation using Deep Learning has become a key topic in image processing and computer vision. A huge variety of algorithms have already been developed for image segmentation, and recently, a substantial amount of works aimed at developing image segmentation approaches using Deep Learning models.

The DeepGlobe Challenge is a challenge where automatic categorization and segmentation is needed for land cover classification. It is a useful challenge for processing satellite imagery, a very powerful data source that currently leads and may lead to many new and exciting applications in the future.

Our approach consists of performing three models for image segmentation: the standard U-Net, UNet combined with VGG16 and the DeepLabV3+ architecture. After implementing these three models we will compare their performance facing this challenge. Our goal is to experiment with different models with the same base architecture and test their performance against each other with a minimal data pre-processing process, rather than to beat the best performing model's accuracy in this field. Our paper consists of the following sections. After the Introduction, in the Literature Review we discuss what have been done previously on this field, how researcher have faced some of the challenges of image segmentation. The Data part describes the Land Cover dataset that we used in detail. In the Methodology section we introduce the characteristics of the models that we implemented. Besides, we describe our image pre-processing approach and the evaluation metrics that we use. In the Results we compare the performance of the different methods. Finally, we conclude our findings.

2 Literature Review

In digital image processing, segmentation is a fundamental stage of the image recognition system. It involves the partitioning of images into multiple segments and objects, (sets of pixels), that allow us to obtain a simpler, more meaningful representation. This representation is easier to analyze, in order to

obtain information for later processes (such as recognition or description). It is normally used to locate limits (lines, curves...) and objects in the images. The result of this process is a set of segments or contours (edge detection) that cover the entire image.

In recent years, Deep Convolutional Neural Networks (DCNN) have achieved breakthroughs in a variety of computer vision tasks in image segmentation and they have been used in several fields such as satellite data, street view, but also medicine. For instance, Ronneberger, Fischer, and Brox (2015), implemented Convolutional Networks (UNet) for biomedical image segmentation of neuronal structures in electron microscopic stacks, improving the training strategy using data augmentation.

This recent advancement in deep learning shows potential and remarkable success in image recognition and spatial resolution applications such as object detection or classification. Land cover information plays an important role in mapping and monitoring changes in the Earth's diverse landscapes and ecosystems Zhang, Han, Han, and Zhu (2020). For example, Volpi and Ferrari (2015) employed semantic segmentation for urban scenes by learning local class interactions. The author proposed modeling the segmentation problem by a discriminatively trained Conditional Random Field (CRF) using Structured Support Vector Machines (SSVM) and improving by 4-6 points the average class accuracy on a challenging dataset involving urban high resolution satellite imagery.

On the other hand, Lin CY (2020) proposed a global and local network architecture (GLNet) incorporating global spatial information and dense local information and reducing segmentation errors. In order to do so, authors modeled the relationship between objects in a scene and also refined the segmentation results using low-level features from the feature map. Their GLNet network architecture achieved 80.8% test accuracy on the Cityscapes test dataset.

In addition, Badrinarayanan, Kendall, and Cipolla (2017) proposed a fully convolutional encoder-decoder architecture for image segmentation. Authors proposed what it's called a SegNet, which is topologically identical to the 13 convolutional layers of the VGG16 network, and a corresponding decoder network followed by a pixel-wise classification layer. The main novelty of SegNet is in the way the decoder upsamples its lower-resolution input feature maps. The decoder network varies between these architectures and is the part which is responsible for producing multi-dimensional features for each pixel for classification.

As we can see there are numerous architectures and their different kinds of implementations for image segmentation. Our focus is going to be on the UNet model. First we implement the standard UNet model, then we complement it with the VGG16 architecture and lastly, we build a DeepLabV3+ model which also roots from the UNet model.

3 Data

The DeepGlobe Land Cover Classification Challenge offers high resolution sub-meter satellite imagery focusing on rural areas. This dataset was collected from the DigitalGlobe Vivid+ dataset.

The whole data file contains 3 different datasets, training(70%), validation(15%) and test(15%) and each set contains 803, 171 and 172 images respectively. This sums up to 1146 images with a size of 2448×2448 pixels. The images contain RGB data with a resolution of 50 cm per pixel, covering 1716,9km$\hat{2}$. In the training data, every satellite image is paired with a mask image for the segmentation. The mask images are RGB, and they differentiate 7 classes, classified in the following table:

Figure 1: ⟨Class Distributions in the Deep Globe Land Cover Classification Dataset ⟩

CLASS	PIXEL COUNT	PROPORTION
URBAN	642.4M	9.35%
AGRICULTURE	3898.0M	56.76%
RANGELAND	701.1M	10.21%
FOREST	944.4M	13.75%
WATER	256.9M	3.74%
BARREN LAND	421.8M	6.14%
UNKNOWN	3.0M	0.04%

1. Urban land: Man-made, built up areas with human artifacts.
2. Agriculture land: Farms, any planned (i.e. regular) plantation, cropland, orchards, vineyards, nurseries, and ornamental horticultural areas; confined feeding operations.
3. Rangeland: Any non-forest, non-farm, green land, grass.
4. Forest land: Any land with at least 20% tree Crown density plus clear cuts.
5. Water: Rivers, oceans, lakes, wetland, ponds.
6. Barren land: Mountain, rock, dessert, beach, land with no vegetation.
7. Unknown: Clouds and others.

In Figure 2 an example of the data is shown:

Figure 2: Training Data Example

(a) Satellite image

(b) Mask/Label image

As we can see the "WATER" parts of the satellite images appear in blue in the mask, the "RANGELAND" parts in pink, the "URBAN" parts in light blue and the "AGRICULTURE" in yellow.

4 Methodology

In this part, we introduce our applied network architectures and then explain each module in detail. We also describe the process of augmentation in this specific case, why it is needed and which techniques we apply. Also, we will discuss possible evaluation metrics that can be used for segmentation problems.

As mentioned before we faced an image segmentation problem. Image segmentation is a method in which a digital image is broken down into various subgroups called image segments, which helps in reducing the complexity of the image, to make further processing or analysis of the image simpler. Segmentation, in easy words, is assigning labels to pixels. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. There are two types of image segmentation. Semantic segmentation is an approach detecting, for every pixel, belonging class of the object. For example, when all people in a figure are segmented as one object and background as one object. For instance, segmentation is an approach that identifies, for every pixel, a belonging instance of the object. It detects each distinct object of interest in the image. For example, when each person in a figure is segmented as an individual object. In our case we had a

semantic segmentation problem.

The encoder-decoder structure is a common architecture of current semantic segmentation algorithms. The structure is composed of an encoder and decoder. Classic image semantic segmentation algorithms such as FCN, UNet, and DeepLab all adopt this structure. The encoder is usually a network (VGG, Resnet, Xcepiton, etc); it consists of a deconvolution layer and upper-sampling layer. Down-sampling is aimed at capturing semantic or context information, while up-sampling is aimed at recovering spatial information. Common decoders include bi-linear interpolation, deconvolution, and dense upsampling convolution.

It is necessary to understand the different operations that are typically used in Convolutional Networks.

1. **Covolution operation:** Which consists of two inputs. The first one is an input of a 3D volume image ($n_{in} * n_{in} * channels$). The second input consists of kernels of size ($f * f * channels$). The output of this operation consists of a 3D volume feature map ($n_{out} * n_{out} * k$).

2. **Max pooling operation:** Its function is to reduce the size of the feature map so that there are fewer parameters in the network.The intuition behind this is to keep only the features which best describe the context of the image from each region.

3. **Transposed Convolution:** Is used to perform an up sampling of the image with learnable parameters. In very simple words, the transposed convolution operation forms the same connectivity as the normal convolution but in the backward direction.

Therefore, both convolution and pooling operations aim to reduce the size of the image, which is better known as **down sampling**. With down sampling the model better understands the content of the image but loses information on the location. Semantic segmentation uses convolutional layers in order to extract features in the encoder and then restores the original size in the decoder in order to classify ever pixel in the original image. The preferred choice to perform up-sampling is Transposed Convolution, which learns parameters through back propagation to convert low resolution images to high resolution images.

As a core approach for multi-class segmentation, we have implemented the UNet architecture that has proven its efficiency in many segmentation problems with limited amount of data, including medical and satellite imaginary tasks.

Besides that, our approach was to use the pre-trained VGG16 model as the encoder portion of a U-Net, and thus, benefit from the already created features in the model, focusing only on learning the specific decoding features. This is called transfer learning, in which pre-trained models are used as a starting point for computer vision and language processing tasks. This is because the development of neural network models for these problems, and due to the enormous leaps in qualifications, requires extensive computing and time resources. The aim of Transfer learning is to improve learning in the target task by using the knowledge from the source task. Transfer learning is an effective technique for reducing training time (Simonyan and Zisserman (2014)). This technique may be related to the development of deep learning models for image classification problems.
As an additional architecture, we trained a DeepLabV3+ model too, which is different from most encoder-decoder designs (for detailed information see Section 4.5).

4.1 Pre-processing: Training and mask generation

We have used the data provided by the challenge in order to train our model. By using the images in the training set

4.2 Pre-processing: Augmentation

Data augmentation is the process that enables to increase the amount of training data by making some reasonable transformations in the existing data. For example, we can augment an image by flipping it vertically or horizontally. We could also rotate it, crop it or even add some noise, which are the most common applications of data augmentation.

These augmentations are usually done in order to reduce overfitting. Data augmentation produces a variety of images with different orientation, location, etc., these images help to increase the variance, therefore making the model more robust. In our case it makes sense to use this as we are dealing with satellite images, therefore some of these transformations are suitable for our data, such as flips.(E.g., example if we were dealing with dogs, the vertical flip would not make sense as in real life we wouldn't find it upside-down on their head). The deep learning method's performance depends on the amount of data. Thus, the data augmentation was performed with keeping the information in the image. The augmentation types that are performed in this paper are the following:

1. Random horizontal flip
2. Random vertical flip

3. Random rotation

We chose these three methods in order to try to achieve a better learning model and therefore, higher accuracy.

4.3 UNet Model. Description and Implementation

The UNet architecture has proven its efficiency in many segmentation problems with data for satellite imaginary tasks and it is one of the most well-recognized image segmentation algorithms. This model was developed by Ronneberger et al. (2015) initially for biomedical image segmentation . Its architecture consists of two paths. The first one is the contraction path (encoder), which is a combination of convolutional and max pooling layers. The second path is a symmetric expanding path which is used to enable localization using transposed convolutions.
In this paper, we have implemented this architecture as one approach for our semantic segmentation task, also known as pixel-based classification, in which we classify each pixel of an image as belonging to a particular class.

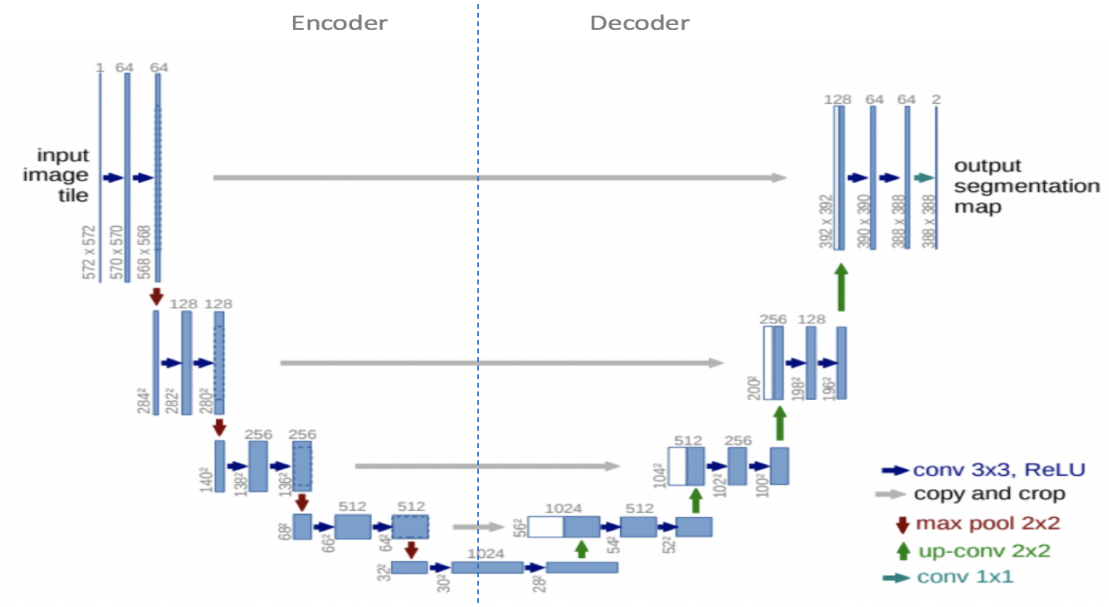
4.3.1 U-Net architecture

Its architecture can be thought as an **encoder** network which is followed by a **decoder** network. Semantic segmentation requires a mechanism to project the discriminative features learnt at the different stages onto the pixel space, in addition to discrimination at pixel level. The following Figure, represents the architecture diagram of this model.

As it can be seen from the diagram the UNet architecture consists of two parts. The first half of this architecture corresponds to the encoder, where the convolution blocks are applied followed by a MaxPool down-sampling. This is done in order to encode the input image and get representations at different levels. On the other hand, the second half of its architecture belongs to the decoder. This part consists of unsampling and concatenation of the higher resolution feature maps from the encoder together with the upsampled features. This allows to better learn representations with the following convolutions Esri (2022).

Semantic segmentation requires discrimination at pixel level, and also a mechanism to project the discriminative features that have been learnt at different stages of the enconder.

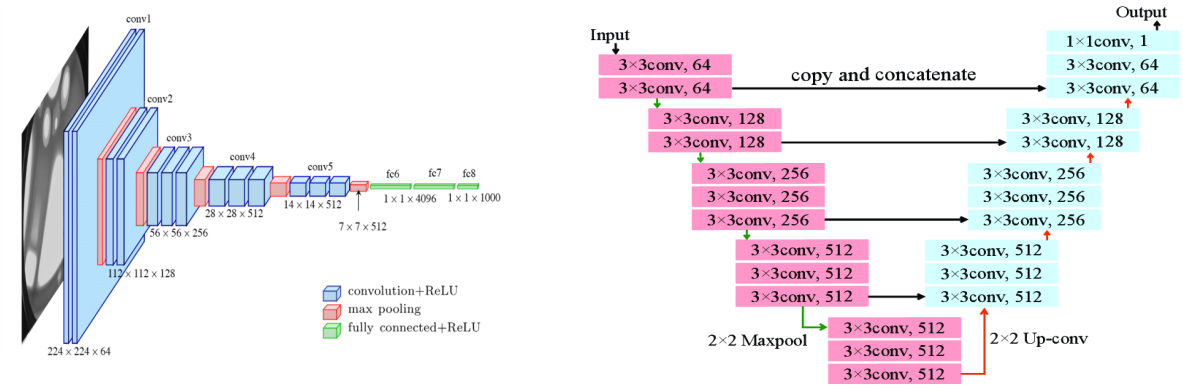
Figure 3: U-Net architecture. The colored arrows represent different operations Esri (2022)



4.4 VGG16

The VGG-net architecture is developed by Simonyan and Zisserman (2014) and has a very deep, yet simple, structure, made up of 16 layers of fully connected convolution. Its simplicity is due to the application of kernels in two layers of convolution and a pooling layer. The authors' main purpose is to show that depth is a vital component of CNN's good performance. This architecture, which has 19 layers, is best known for its pyramid-like shape in which the layers closer to the image are wider, and the deeper layers are deeper. The VGG-Net neural network is illustrated in Fig. 4. It contains a series of convolutional layers, behind which there are pooling layers that make the layers smaller. The VGG network is presented in two architectures, called VGG16 and VGG19.

Figure 4: VGG16 and UNet



(a) Standard VGG16 architecture

(b) UNet with VGG16 architecture

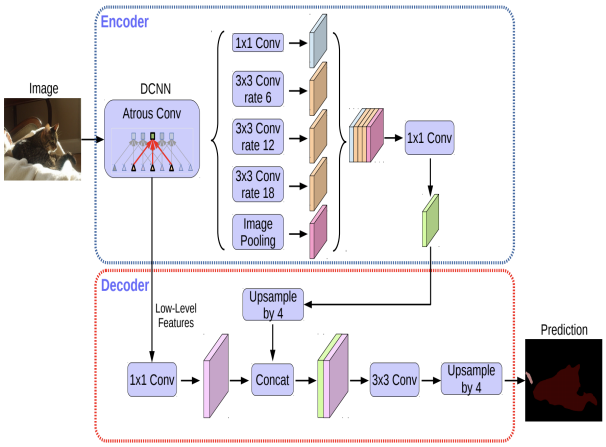
4.5 DeepLabV3+

Deeplab offers a different approach to semantic segmentation. It presents an architecture for controlling signal decimation and learning multi-scale contextual features. Deeplab was developed by **?** who combined atrous convolution, spatial pyramid pooling and fully connected CRFs. It uses an ImageNet pre-trained ResNet as its main feature extractor network. However, it proposes a new Residual block for multi-scale feature learning. Instead of regular convolutions, the last ResNet block uses atrous convolutions. Also, each convolution (within this new block) uses different dilation rates to capture multi-scale context. Additionally, on top of this new block, it uses Atrous Spatial Pyramid Pooling (ASPP). ASPP uses dilated convolutions with different rates as an attempt of classifying regions of an arbitrary scaleChen, Papandreou, Schroff, and Adam (2017a).

DeeplabV3 is an improved version of DeepLab. Chen, Papandreou, Schroff, and Adam (2017b) have revisited the DeepLab framework to create DeepLabv3 combining cascaded and parallel modules of Atrous convolutions. The authors have modified the ResNet architecture to keep high resolution feature maps in deep blocks using Atrous convolutions.

? finally released the Deeplabv3+ framework using an encoder-decoder structure. The authors have introduced the Aatrous separable convolution composed of a depth-wise convolution (spatial convolution for each channel of the input) and point-wise convolution (1x1 convolution with the depth-wise convolution as input).

Figure 5: DeepLabV3+ architecture



4.6 Evaluation metrics

Evaluating the quality of semantic segmentation is an important process in image processing. Besides the classical classification metrics, such as Accuracy or Recall, image segmentation tasks require a slightly different approach. There are three essential metrics, that are worth to be taken into consideration when evaluating an image segmentation task.

- Pixel Accuracy

  Pixel accuracy is perhaps the easiest to understand conceptually. It is the percent of pixels in your image that are classified correctly. However it is not the best metric. Because in its case high pixel accuracy does not always imply superior segmentation ability. In case of class imbalance, pixel accuracy performs really bad. Class imbalance means that a class or some classes dominate the image, while some other classes make up only a small portion of the image, which can lead to a high accuracy score, which comes from the basic fact that the predicted class is the majority class.

- Intersection-over-Union

  The Intersection over Union (IoU) metric, also referred to as the Jaccard index (Jaccard (1912)), is essentially a method to quantify the percent overlap between the target mask and our prediction output. This metric is closely related to the Dice coefficient which is often used as a loss function during training. Quite simply, the IoU metric measures the number of pixels common between the target and prediction masks divided by the total number of pixels present across both masks.

$$IoU = \frac{target \cap prediction}{target \cup prediction}$$

The intersection ($A \cap B$) is comprised of the pixels found in both the prediction mask and the ground truth mask, whereas the union ($A \cup B$) is simply comprised of all pixels found in either the prediction or target mask. The IoU score is calculated for each class separately and then averaged over all classes to provide a global, mean IoU score of our semantic segmentation prediction. It is the most common evaluation metric to use for image segmentation, nonetheless it gives a better accuracy score to the overall model than a "simple" accuracy regarding the fact that is was developed specifically to measure the performance of image segmentation models.

- Dice Coefficient

  Dice coefficient is very similar to Jaccard's Index. Dice coefficient double counts the intersection (True Positives).

  $$Dicecoefficient = \frac{2 * (target \cap prediction)}{target \cup prediction} = \frac{2 * TruePositives}{2 * TruePositives + FalseNegatives + FalsePositives}$$

  We apply the Dice coefficient as our loss function for the VGG16 and DeepLab models.

5 Results

Our goal was to compare the performance of different architectures built for image segmentation. We compare our models along four metrics. Besides, the Intersection Over Union evaluation metric we included three additional metrics to evaluate the performance of our models. These are the Categorical Accuracy (it is a built-in function in Keras, it calculates how often predictions match one-hot labels), and some classical classification measures, such as the Recall (also called as sensitivity, true positives compared to the amount of true positives and false negatives) and the Precision (also called as positive predictive value, it is the fraction of true positives vs the amount of true positives and false positives). Furthermore, we decided not to use the Dice Coefficient as an evaluation metrics given the fact that it basically measures the same thing as IoU and including this measure, would not have an added value. We evaluated both the train and the test datasets with the validation data. The overall results of the evaluation metrics are shown in Table 1 and 2. We also include the actual prediction results of one, randomly chosen image from the test data. This can be seen on Figure 6.

As the results show, besides the UNet model we could not achieve relatively high accuracy. We compare our results to the average of many other studies that used this dataset and experimented with different

Table 1: In-sample (train data) evaluation

model	Categorical Accuracy	Recall	Precision	IoU
UNet	57.89%	91.83%	20.63%	24.13%
UNet+VGG16	10.94%	-	-	44.05%
DeepLabV3+	57.89%	58.77%	23.91%	15.2%

Table 2: Out-of-sample (test data) evaluation

model	Categorical Accuracy	Recall	Precision	IoU
UNet	57.89%	20.64%	91.83%	19.07%
UNet+VGG16	10.84%	-	-	44.05%
DeepLabV3+	57.89%	58.87%	24.2%	15.2%

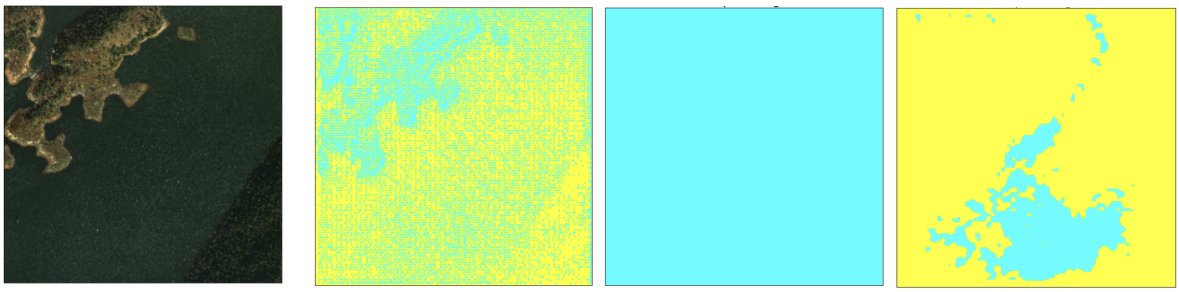
architectures in order to improve classification accuracy. We suspect that this effect is due to the properties of the dataset: it appears that the training and validation sets were sampled from different regions with different appearances. In particular, we have noticed that forests inside each dataset are similar to each other but the training set has almost exclusively coniferous forests while the validation set has almost exclusively greenwood.

Surprisingly the IoU measure of the UNet+VGG16 implementation is the highest, however this model performs the worst from every other aspect. Especially if we look at its predicted image, we can clearly see that this model was not able to capture any of the attributes of the images. During training the DeepLabV3+ we also noticed that the higher the probability that an image is augmentationed, the higher the IoU score is, which is quite surprising for us. In general image augmentation improves accuracy, but in our case we expected that augmentation would not significantly contribute to better accuracy. Mainly because for these kind of images only a few types of image transformation techniques are truly meaningful. However the DeepLabV3+ also performed quite well, the Category Accuracy and the Precision measures are very close (or the same as) to the standard UNet's results.

The results obtained with the UNet model are quite surprising as the accuracy obtained with this model is higher than expected, compared to the results from the UNet + VGG16 model.

Nevertheless, as shown in the figure and in the tables presented the overall performance is not as good as we would have wished and the prediction is not as accurate as we hoped for.

Figure 6: Test Data Prediction Examples



(a) Original image (b) UNet (c) UNet+VGG16 (d) DeepLabV3+

6 Discussion and Concluding Remarks

In this work, we have presented an approach to land cover classification for satellite imagery based with the standard UNet architecture. We implemented 3 models and compared their results. The standard UNet model; the standard UNet model with VGG16 as encoder and a DeepLabV3+ model design using ImageNet as a backbone model with a Atrous Spatial Pyramid Pooling (ASPP) block and Batch Normalization. We used three different data augmentation types in the data generator with an initial 0.1 probability of an image being transformed by each augmentation type. We analyzed our trained models with different performance metrics such as the IoU or Categorical Accuracy.

Based on the results we achieved and the metrics presented in Section 5, we can conclude that the overall performance of the standard UNet is the one that presents the most satisfying results, yet still not quite as accurate as we would have hoped. On the other hand we find that the UNet+VGG16 has the highest IoU out of the three models, even though its categorical accuracy is the lowest and its actual predictions of the images are usually just one-color images. Our findings are limited due to computational, time and data constraints, and they leave a lot of room for improvement. For future work, it could be interesting to try to improve the models by changing some of the parameters of the defined functions and try to achieve better accuracy and performance overall.

References

Badrinarayanan, V., A. Kendall, and R. Cipolla (2017): "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 2481–2495.

Chen, L., G. Papandreou, F. Schroff, and H. Adam (2017a): "Rethinking Atrous Convolution for Semantic Image Segmentation," *CoRR*, abs/1706.05587.

Chen, L.-C., G. Papandreou, F. Schroff, and H. Adam (2017b): "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*.

Esri (2022): "How U-net works?" .

Jaccard, P. (1912): "The distribution of the flora in the alpine zone. 1," *New phytologist*, 11, 37–50.

Lin CY, Chiu YC, N. H. S. T. L. K. (2020): "Global-and-Local Context Network for Semantic Segmentation of Street View Images." .

Ronneberger, O., P. Fischer, and T. Brox (2015): "U-Net: Convolutional Networks for Biomedical Image Segmentation," .

Simonyan, K. and A. Zisserman (2014): "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*.

Volpi, M. and V. Ferrari (2015): "Semantic segmentation of urban scenes by learning local class interactions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7-12, 2015*, IEEE, 1–9.

Zhang, X., L. Han, L. Han, and L. Zhu (2020): "How Well Do Deep Learning-Based Methods for Land Cover Classification and Object Detection Perform on High Resolution Remote Sensing Imagery?" *Remote Sensing*, 12.