**Unet Description**

**Ensembling of model**

**Problem Formulation**

Rapidly changing environment is mainly related due to the deforestation of the land due to various purposes by the humans. Global warming and effected bio-diversity are two major prominent effects of deforestation problems but there are other’s as well such as desertification, flooding, soil erosion, increased greenhouse gases, and fewer crops etc [1]. In order to retore the environment forestation of the land is very important. Autonomous seeding of the deforested land has appeared to be a potential solution for effective forestation of available land. Seeding requires specific land with certain specific conditions to be successful. UAVs can be a potential solution for autonomously seeding the land for effective forestation.

It is important for autonomous seeding agents such as UAVs to autonomously detect the potential spots for seeding with high rates of success. The main underlying problem is that how a UAV find a potential spot for seeding. A UAV collect aerial view of the terrain throughout the flight so there should be a way for a UAV to distinguish a potential spot within a generic aerial scenery of the area. A major problem can be locating and identify area of interest in aerial images under different conditions such as uneven illumination, low levels of contrast, or partially restricted visibility of the potential land in images. Aerial images of forest are very dense and there are marginal apparent differences between forest and available land due to other greenery in the environment, so it becomes a difficult problem to distinguish among a different part of forest i.e., forest and available land as shown in figure 1 and 2.

 A picture containing plant

Description automatically generated

**Forest Region**

Figure 1 Sample Image with non easily distinguishable regions

**Non-Forest Region Region**

**Non-Forest Region Region**

**Forest Region**

Figure 2 Sample Image with Easily distinguishable regions

**Description of dataset**

# Methodology

The adopted methodology involves semantic segmentation of the aerial images of forest using deep learning techniques using ensembling of 3 different U-Net architectures with different backbones. The implementation consists of 4 major steps as shown in Figure 1. 1st step consists of loading the training images along with their labelled masks and resize to cope with limited memory of the system. 2rd step is learning a good deep model capable of effective pixelwise classification of an image. In in last step trained model is then evaluated for accuracy of detection as well as localization of regions belonging to the non-forest areas in the images.

U-net Architecture

## Data Pre-processing

The dataset used for model training and evaluation consists of 5108 aerial images of dimensions 256x256. The dataset is obtained from land cover classification track in DeepGlobe Challenge (site the dataset). Each image has been provided with a binary mask image as label for the image the defines region of interest to distinguish between forest and non-forest regions in the image.

**Table 1 Description of dataset**

|  |  |
| --- | --- |
| Data | 5108 labelled images |
| Resolution | 256x256 |

## Model learning

### Model architecture

Model trained for solving the problem at hand consists of U-net with 3 different backbones i.e. VGG19, ResNet and InceptionV3. Unet itself is an extended implementation of an encoder-decoder topology of fully convolutional network. The intuition behind Unet is to encode the image through a CNN (encoder-downsampling stage) and then decode it back (decoder-upsampling stage) that gives the segmentation mask. The backbone is the architectural element which modifies the layers in the encoder network and hence modifies the feature extraction and encoding approach area and hence it also determines how the decoder network should be built accordingly that compliments the new modified encoder network.

Since each architecture among Vgg, ResNet and Inception excels in different application hence combining all three provides best results in our implementation. Figure 2 shows a general architecture of U-Net implemented with ResNet backbone. All the three models have been

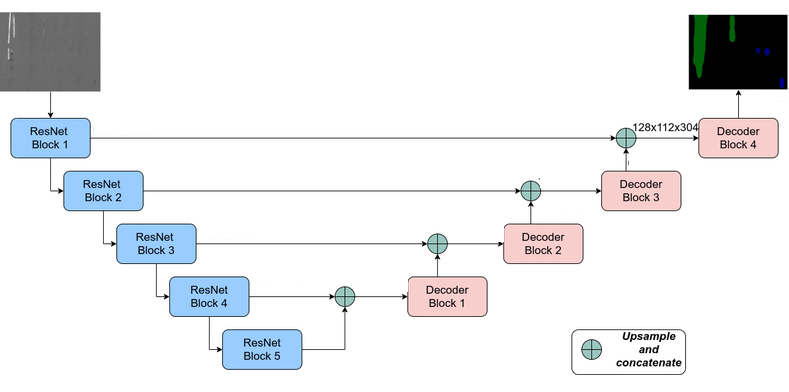


Figure 3 U-Net with ResNet Backbone architecture

A picture containing text, clipart

Description automatically generated

Figure 4 ResNet block ([3]"Understanding and visualizing ResNets", Medium, 2022. [Online]. Available: https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8. [Accessed: 24- Feb- 2022].)

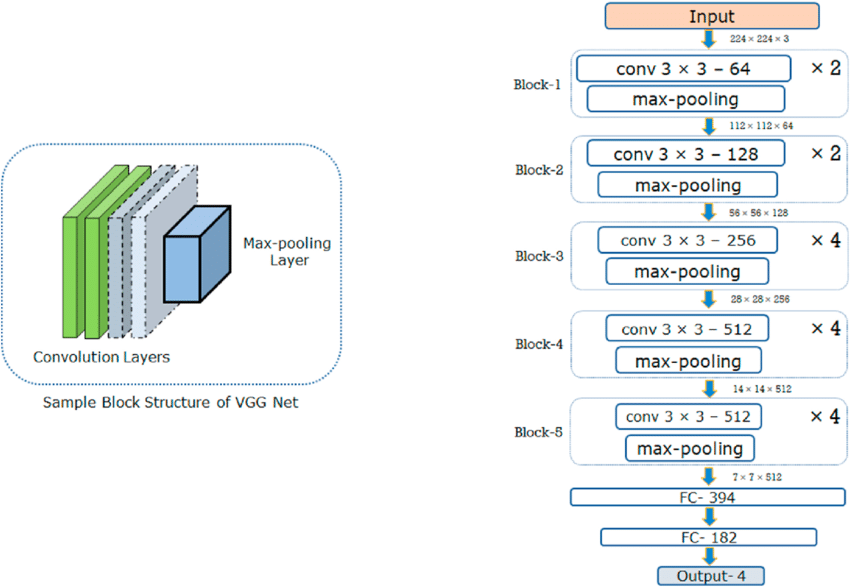


Figure 5 Unit Block of Vgg19 Network [[2]2022. [Online]. Available: https://www.researchgate.net/figure/Network-architecture-of-finetuned-VGG19-a-Sample-Block-structure-of-VGG-Net-b\_fig5\_351921164. [Accessed: 24- Feb- 2022].]

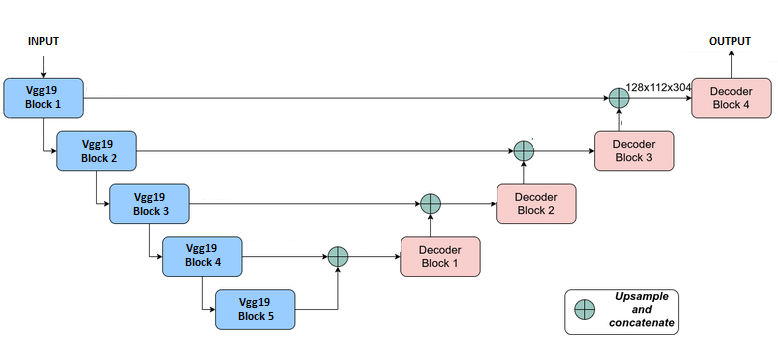


Figure 6 U-Net with Vgg19 backbone acrhitecture

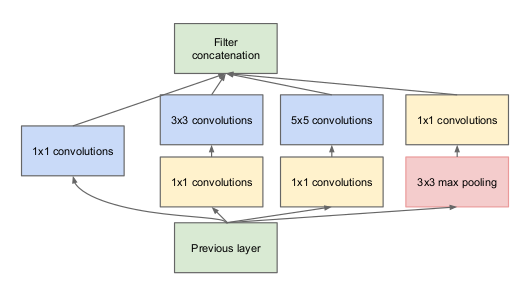


Figure 7 Basic arcitecture block of InceptionV3 ([3]"Understanding and visualizing ResNets", Medium, 2022. [Online]. Available: https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8. [Accessed: 24- Feb- 2022].)

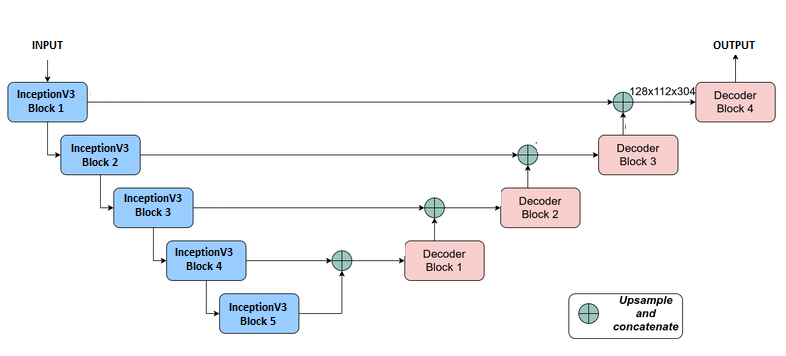


Figure 8 U-Net with ResNet backbone acrhitecture

Summary of all three models is given in appendix.

All the models are initialized with pretrained weights on ImageNet which not only reduces the training time but improves the learning capability model in small dataset.

**Ensembling approach**

All the 3 different variants of U-net are ensembled together to give pixel level classification of each image.

Random weights selected are

w1 =0.3 , w2= 0.5 , w3= 0.2

A best combination of weights for each model is selected that gives best result on prediction based on IoU score of predicted mask.

Final out

3 maske

A combination of weighted average of model gives the final prediction

P= w1\*p1 + w2\*p2 + w3\*p3

Where

P= final prediction

P1 = prediction by model 1

P2 = prediction by model 2

P3 = prediction by model 3

**Evaluation strategy**

The presented approach is evaluated for its accuracy of distinguishing forest regions from non-forest regions in the aerial images. The model tested on already separated test from original dataset consists of 500 images. At test time the implemented model produces for given image a binary image with two regions defined as forest and non-forest images. These predicted binary images are then compared with the provided labels to calculate pixel wise results. IoU factor is used for evaluating the predicted region with provided one.

IoU is measured using following mathematical relation:

IoU scores can further be converted to mean average precision, but further analysis is not performed due to time constraints.

Since the data was poorly label already as can be seen in the figure 4 hence IoU score did not provided got score despite the fact visual inspection showed extremely impressive prediction of the forest and non-forest regions as shown in Figure 9 -10.

Chart

Description automatically generated

Figure 9 Prediction with Bad IoU scores

Chart

Description automatically generated

Figure 10 Predictions with Good IoU scores

**Results and Discussion**

The ensembling approach presented in above section provided an average pixel classification accuracy of 32%. This low value of accuracy is due to poor labeling of the dataset. Visual inspection of the predicted images has shown better results on average compared to the provided labels and hence despite the low accuracy score the trained model gives acceptable results for real application.

The main rule has been played by the weight initialization using ImageNet trained weights. The models with weights trained on ImageNet has already learned the notion of learning useful object in the real world images. Training on aerial images further allowed the model to adapt according to the required image regions.

**References**

[1]"Effects of Deforestation | The Pachamama Alliance", 2022. [Online]. Available: https://www.pachamama.org/effects-of-deforestation. [Accessed: 24- Feb- 2022].