Deep Learning for Image Segmentation of On-Road and Off-Road Navigation Data Set.

Dinesh Adhithya

December 6, 2021

1 Introduction

This work uses Convolutional neural network models to segment images into masks where various components such as vehicle , road , sky , etc are identified and could be used to build autonomous navigation systems.We use data sets containing images and their corresponding masks(segmented images where each pixel is labelled belonging to a class).The various classes include: sky,road,vehicle,bicycle,etc.

This work makes use of the RUGD[1] data set which was built for off-road autonomous navigation and understanding semantic understanding of outdoor environments.The data set is comprised of video sequences captured from the camera on board a mobile robot platform.For on road navigation we use the CAMVID[2] [3] data set captured through fixed CCTV style cameras with a perspective of a driving automobile.

1.1 Semantic Segmentation

Involves task assigning each pixel in an image to a particular class , this class can be various objects relevant to the problem at place.

The on-road image segmentation has the following 12 classes : Sky, Building, Column Pole, Road, Sidewalk, VegetationMisc, Traffic Light, Fence, Vehicle, Pedestrian, Bicyclist, Void

The off-road image segmentation has 4 classes: Traversable, Non-Traversable, Sky and obstacles.

1.2 Deep learning for image segmentation

Multiple image segmentation algorithms have been developed. Earlier methods include thresholding, histogram-based bundling, region growing, k-means clustering, or watersheds. However, more advanced algorithms are based on active contours, graph cuts, conditional and Markov random fields, and sparsity-based methods.
Over the last years, Deep Learning models have introduced a new segment of image segmentation models with remarkable performance improvements. Deep Learning based image segmentation models often achieve the best accuracy rates on popular benchmarks, resulting in a paradigm shift in the field.

1.3 UNET Architecture

UNET[4] is a fully connected convolutional neural network designed for image segmentation task.The network has an u-shape structure to it. The contracting path (encoder) consists of convolutional networks followed by ReLU activation function , then max pooling operation. During the contraction, the spatial information is reduced while feature information is increased. The expansive pathway

combines the feature and spatial information through a sequence of up-convolutions and concatenations with high-resolution features from the contracting path.

1.3.1 Metrics for Evaluation

In this work we will use accuracy as the metric to train our models , with least squares error as quantify loss.We will also use intersection to union ratio and dice score to measure model performance.

Intersection of Union = Area of overlap / Area of union

Dice Score = 2* Area of overlap / Sum of Area

2 On road Semantic Segmentation

2.1 Deep Learning Architecture

We make use of Encoder-Decoder CNN (Convolutional Neural Network) architecture for image segmentation.For the Encoder component we use the VGG16 architecture with transfer learning techniques used for model training. The Decoder component UNET architecture was used , which has been well established architecture for image segmentation tasks.

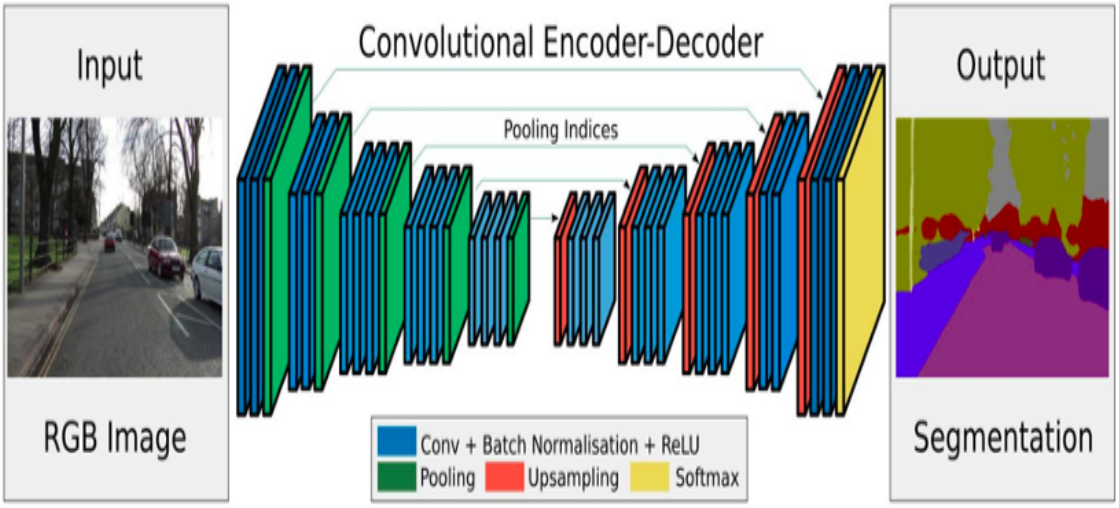

Figure 1: Image showing a encoder-decoder architecture for image segmentation.Link of the image above.

This work we use a VGG16-UNET model for image segmentation. The below images show the model architecture.

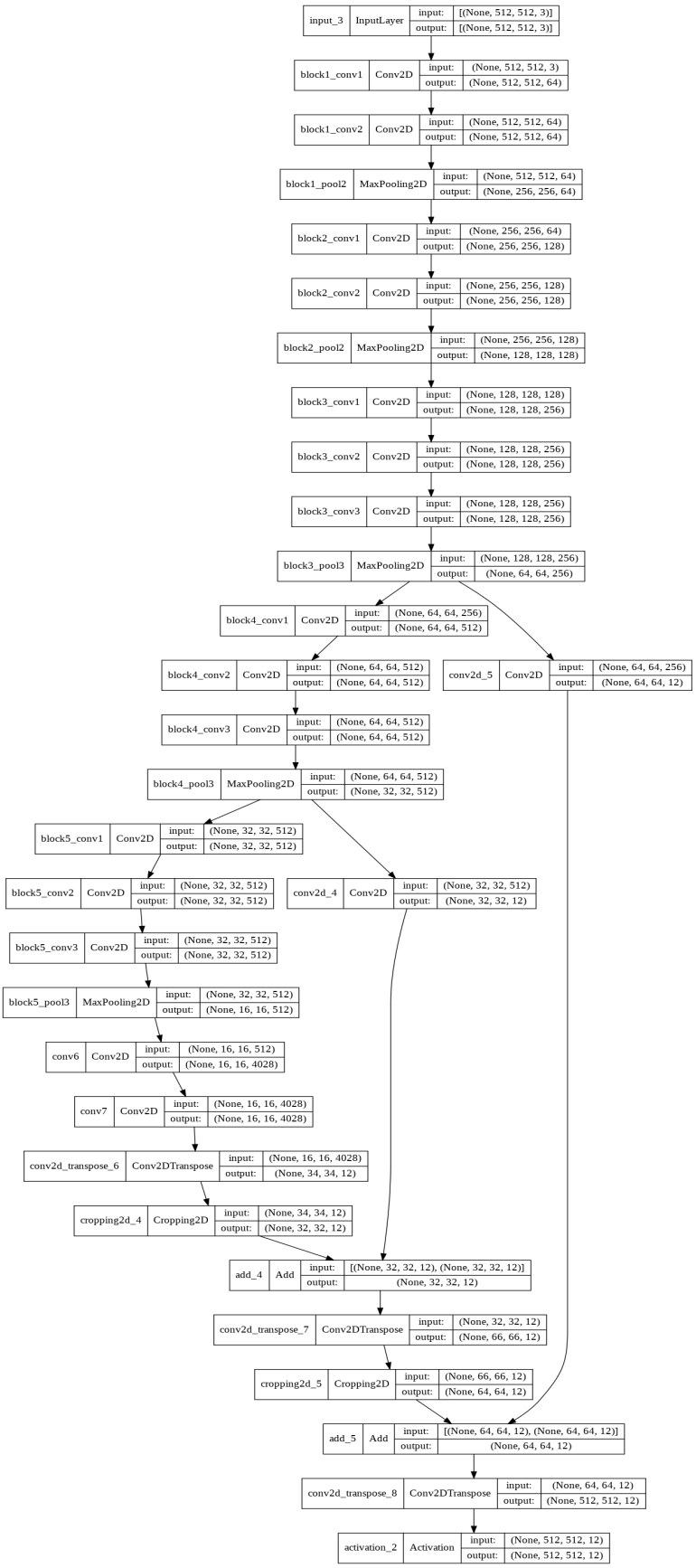

Figure 2: Image showing Deep learning architecture used in this model implementation in tensorflow.

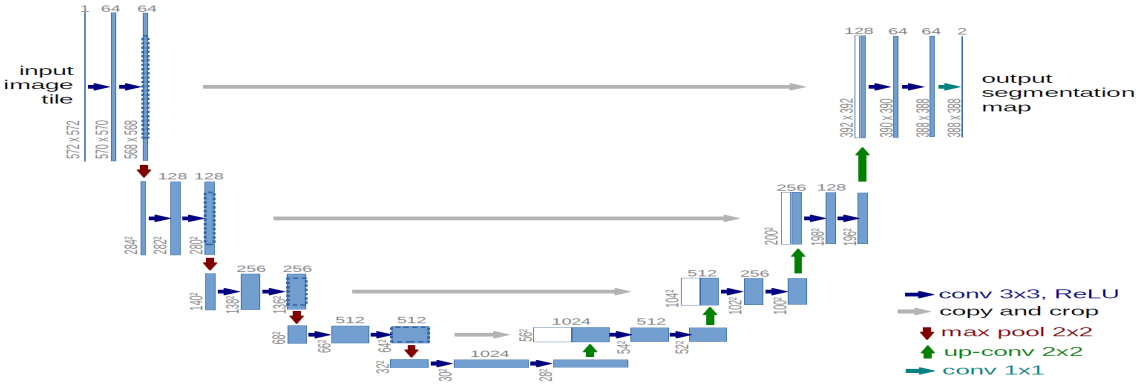

Figure 3: Image showing a VGG16-UNET architecture for image segmentation.Link of the image above.

2.2 Image Preprocessing

The images were augmented using various transformation randomly on images such as image rotation , image brightness changes , grid and optical distortion. These images were then stored. From the CAMVID data set we select 367 images for training and 101 images for testing the trained model.We preprocess images by resizing them to 512*512 size before passing onto the model.

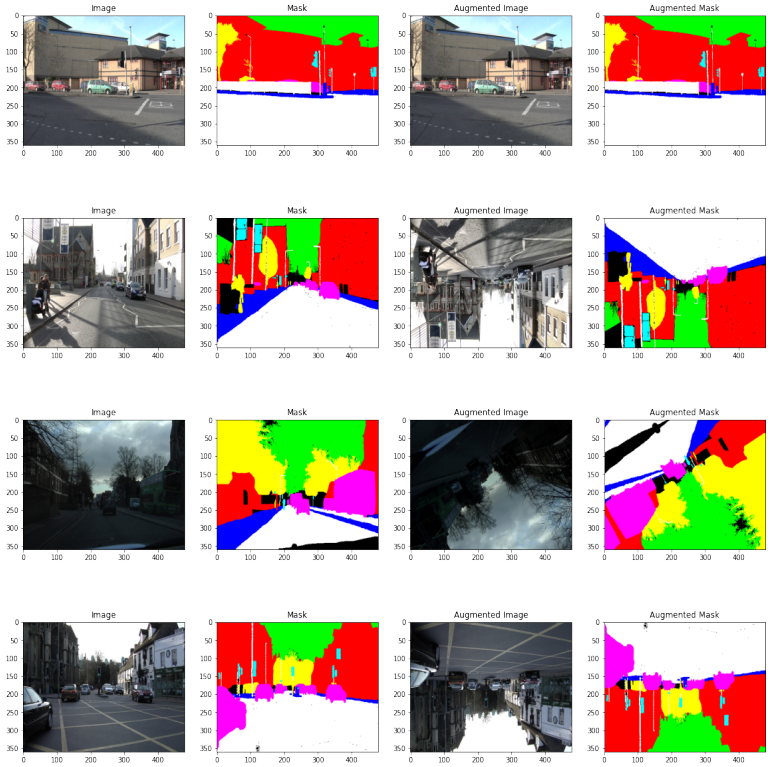

Figure 4: Images and Masks obtained from data set after augmentation.

2.3 Model Training

The model was compiled using a stochastic gradient descent solver. The mean squared error function is used as loss function and to quantify the model performance its accuracy is used as the metric.

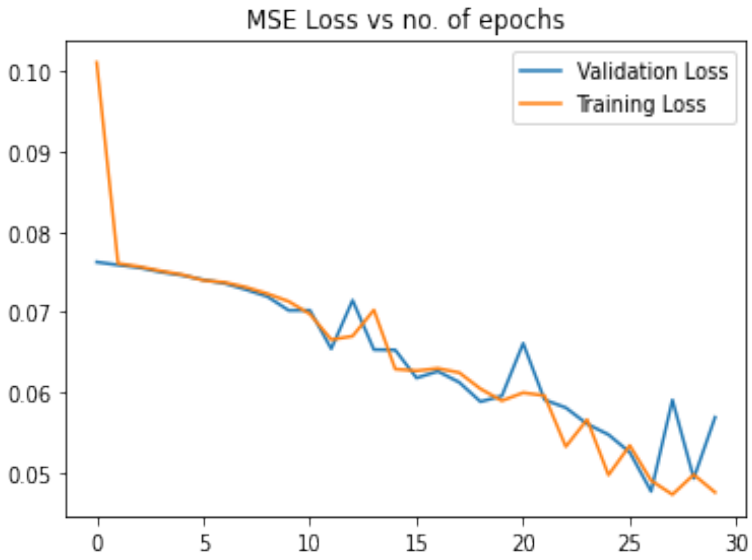

Figure 5: Image showing model's loss versus no. of epochs.

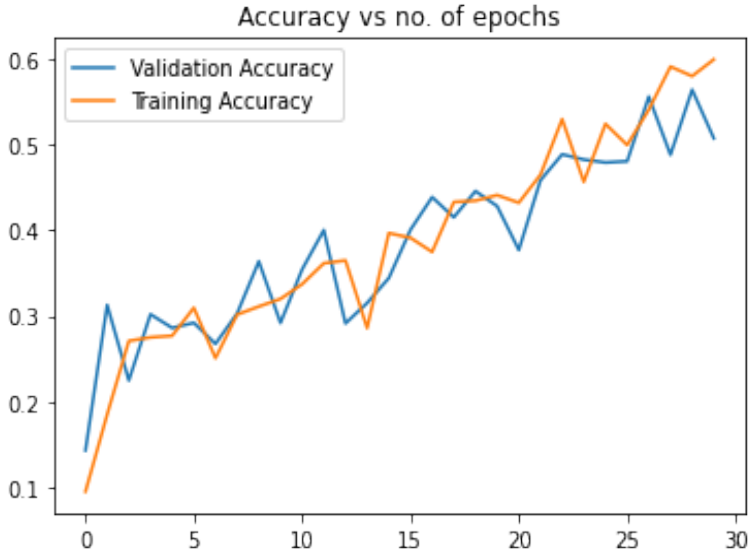

Figure 6: Image showing model's accuracy versus no. of epochs.

2.3.1 Model Performance

The model achieves a testing data accuracy of 0.61 and training accuracy of 0.63.The images of expected masks and masks actually obtained are plotted below. We can see that the model prediction aren't good for building autonomous navigation systems.Never the less its predictions capture at least the road, building and sky objects.

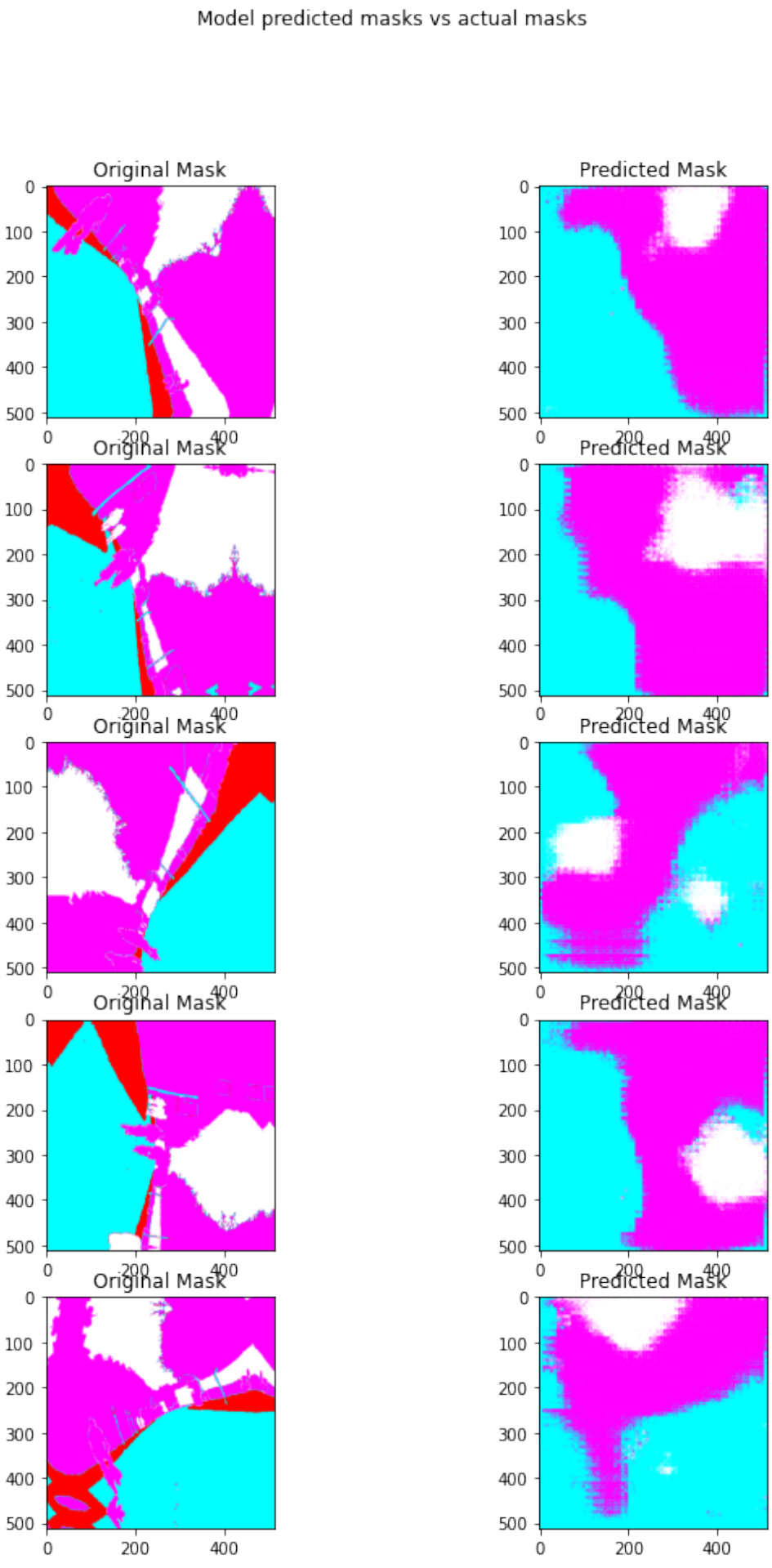

Figure 7: Predicted Masks vs Actual Masks.

2.3.2 Visualization Outputs at Various Model layers

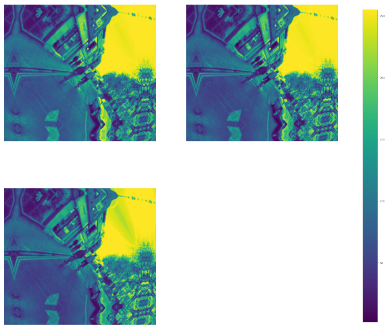

Figure 8: Image at Input Layer

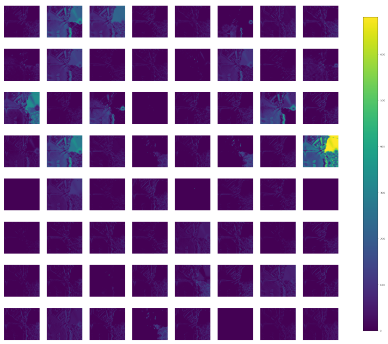

Figure 9: Image at encoder segment of the model.

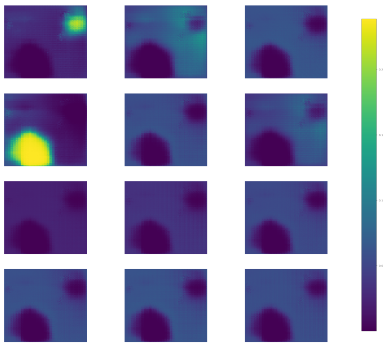

Figure 10: Images at encoder segment of the model

3 Off road semantic segmentation

3.1 Model Architecture

This model is also an UNET encoder-decoder model for semantic segmentation of off road image segmentation data available through the RUGD data set.

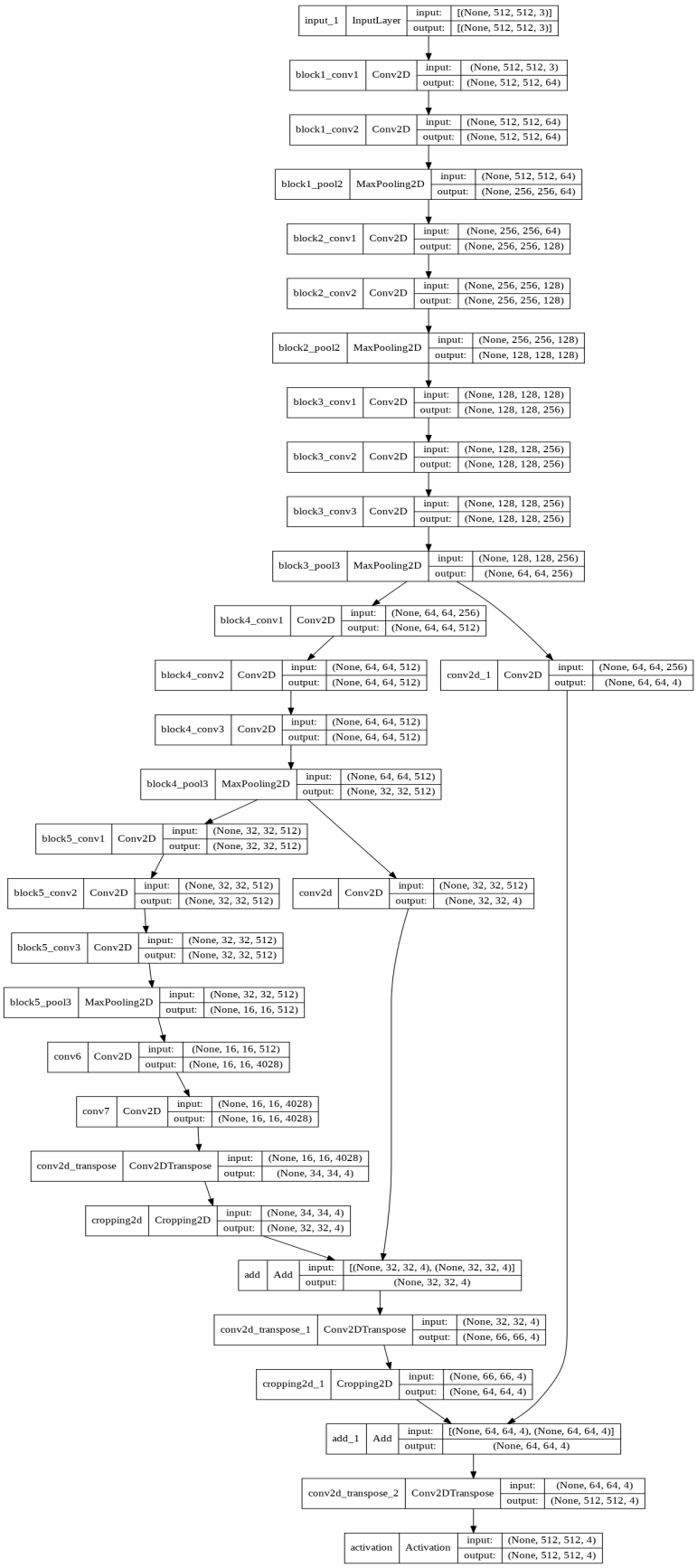

Figure 11: UNET Model used segmentation of RUGD data set.

3.2 Image Processing

There are 801 training images and masks and 200 images for testing our model's performance. The images were resized to 512*512 to fit the model's requirements then various functions are applied for augmenting the data set such as image rotation , vertical flip, horizontal flip , image grid distortion and brightness distortion.

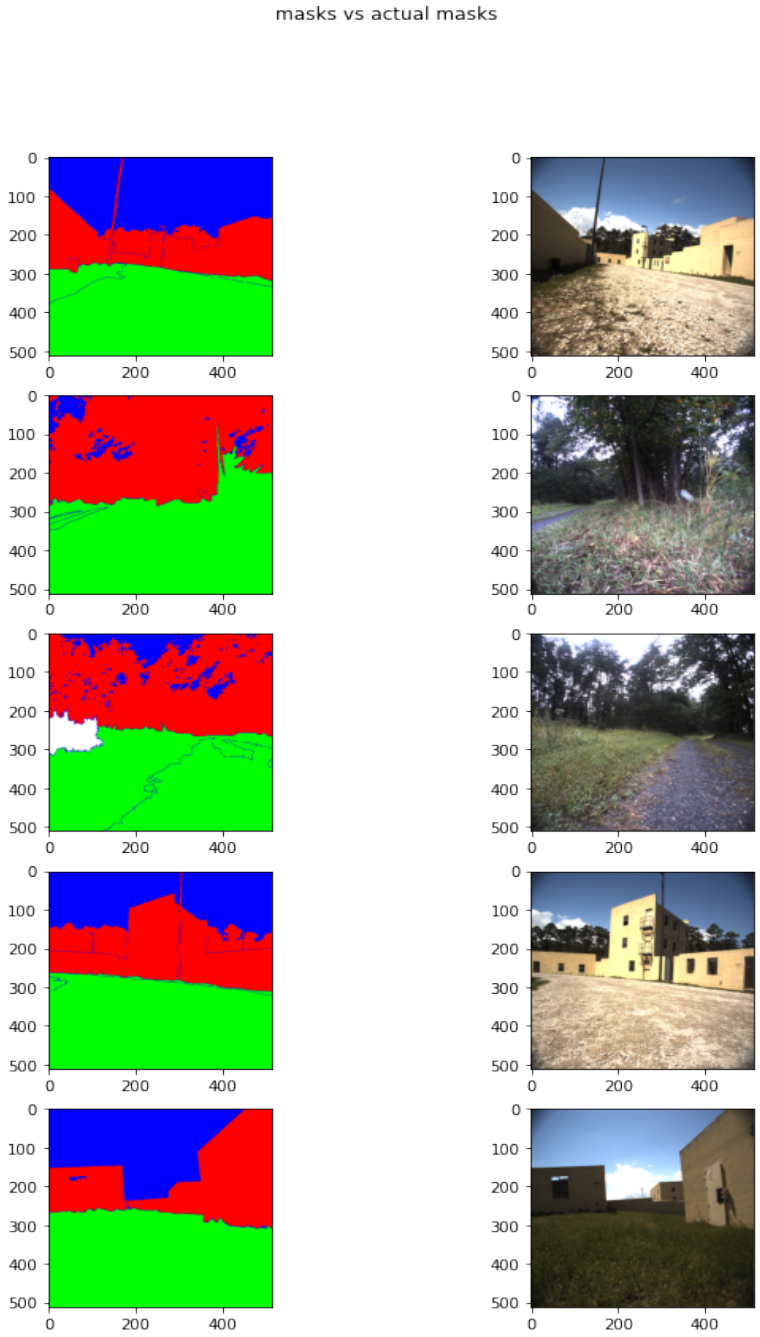

Figure 12: Images and Masks from the RUGD Data set.

3.3 Model Training

The model was compiled using a stochastic gradient descent solver. The Categorical cross entropy function is used as loss function and to quantify the model performance its accuracy is used as the metric.

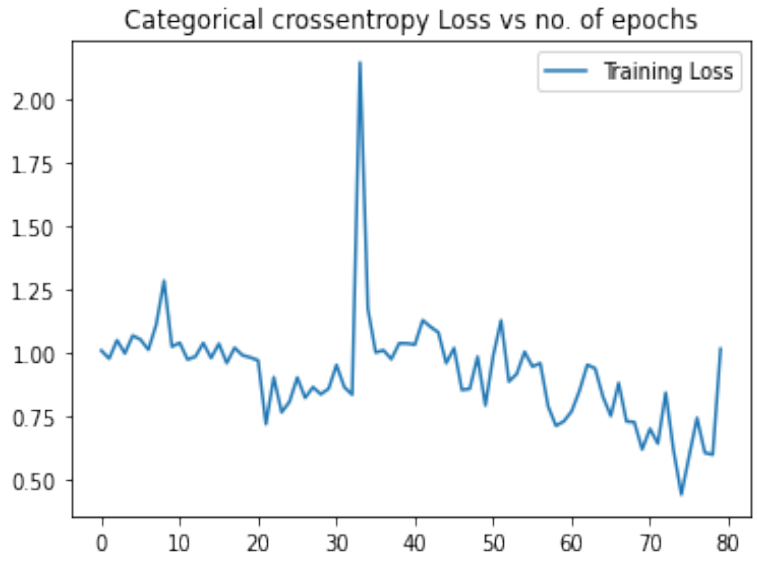

Figure 13: Image showing model's loss versus no. of epochs.

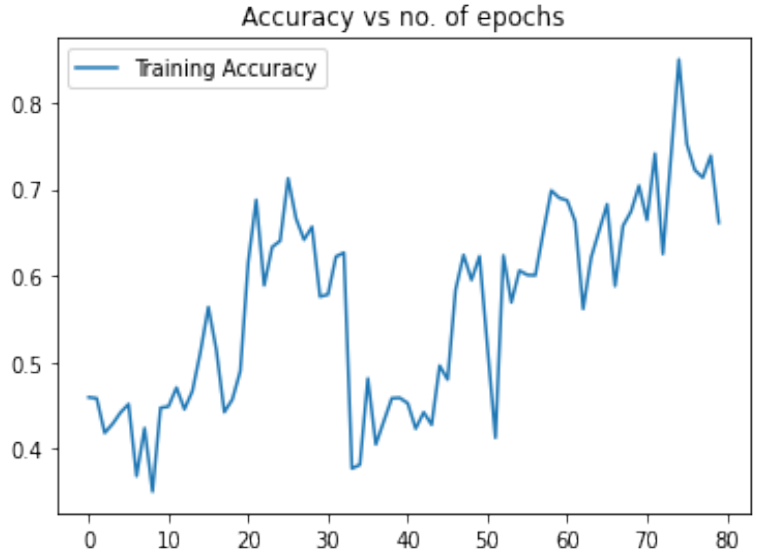

Figure 14: Image showing model's accuracy versus no. of epochs.

3.4 Model Performance

The Mean IOU and Dice score were used as metrics to quantify the model's performance.

Table 1: Table showing performance model on RUGD Dataset

Metric	Sky	Traversable	Non-Traversable	Obstacles
Mean IOU	0.64	1.0	1.0	1.0
Dice score	0.64	1.0	1.0	1.0

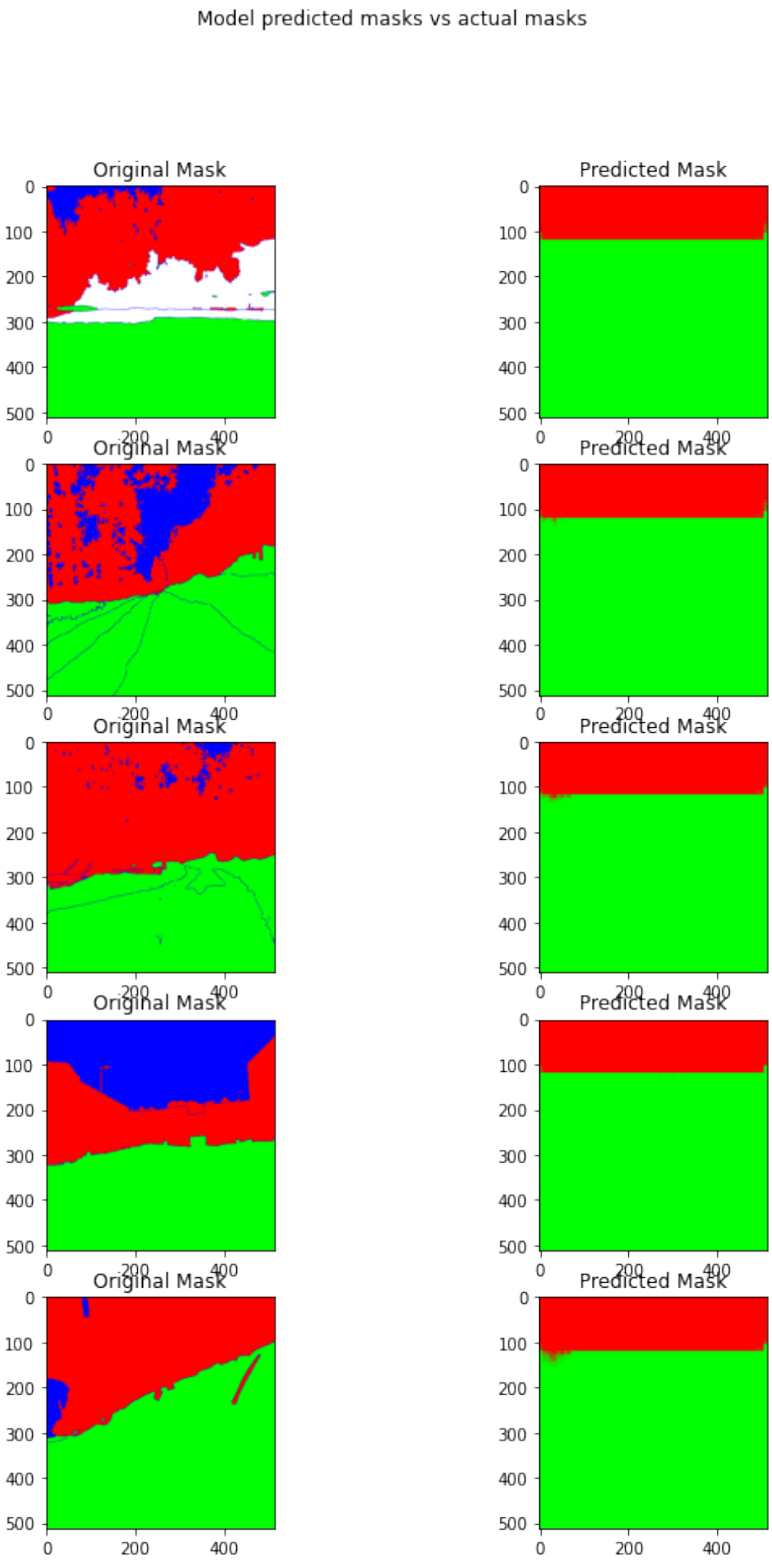

Figure 15: Predicted Masks vs Actual Masks.

4 Discussions and Conclusion

Both the models suffer from over fitting to classes present more in number and perform meagerly at 60 % accuracy. The UNET architecture although can be used for image segmentation as it shows certain capabilities for semantic segmentation tasks and in the future models such as BiSeNet-V2 and HRNETV2 can be used , but they also require more GPU capabilities which isn't quite supported by Google Colab. These are far more proven to perform image segmentation tasks more accurately and these models can be used for future work.

The above models were implemented in jupyter notebooks [5] using the Tensorflow package [6] in python language. Stack exchange has also served as a good resource for debugging Latex and Python scripts. All the codes used for this work can be found at the GitHub repository. Any changes you wish to suggest can be reported on GitHub itself.

References

[1] Maggie Wigness, Sungmin Eum, John G Rogers, David Han, and Heesung Kwon. A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments. In *International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[2] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx, 2008.

[3] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57, 2008.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015.

[5] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows.

[6] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.