National College of Ireland

Performance Analysis of Convolution Neural Networks Using Semantic Segmentation for Driving Scenes

MSc Research Project
MSc Data Analytics

Vishal Kumar Yadav
Student ID: 19239236

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vishal Kumar Yadav
Student ID:	x19236239
Programme:	MSc Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	16/08/2021
Project Title:	Performance Analysis of Convolution Neural Networks Using Semantic Segmentation for Driving Scenes
Word Count:	7500
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vishal Kumar Yadav
Date:	23rd September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Performance Analysis of Convolution Neural Networks Using Semantic Segmentation for Driving Scenes

Vishal Kumar Yadav
x19236239

Abstract

Since past decade many real time system have been developed. Technologies such as autonomous vehicles, virtual reality systems, drones have been using application of machine learning techniques to perform there task. Thus involves observation, planning, as well as execution in ever-changing situation, safety and accuracy are important factors. The focus of this study is to segment images by utilizing deep learning techniques in order to aid in better understanding of driving scene perception in order to help autonomous driving systems in distinguishing between ground reality and prediction. Deep learning models such U-Net, FCN and FPN are best among state-of-art method that are used for providing solutions to real world problems. For this research U-Net architecture and FPN architecture were used on camvid driving scene dataset and image segmentation was performed to understand the driving scenes to analyse the difference in ground reality and machine result. Both models showed good results in comparison to other methods. U-Net model is applied with ResNet-50 and FPN with ResNeXt networks for segmentation on driving scene, where both models were evaluated on basis of IoU and Dice loss. U-Net achieved IoU score of 82% and FPN architecture model achieved IoU score of 84%.

Keywords: Deep Learning, Image Segmentation, U-NET, FPN, RESNET.

1 Introduction

1.1 Overview

The term "image semantic segmentation" refers to the process of giving a preset name to each pixel in an image that indicates its semantic category. Image semantic segmentation is now a critical topic in image analysis and machine learning development. It is indeed utilized in a variety of disciplines, including medical image analysis, automated driving, video recognition, and others(Yi; 2021)

In the bigger picture, image segmentation is important tasks that provides deep insight about the events in the scene. Due to accidental human errors, human driving is still troublesome, resulting in both human and economic losses. Researchers have proposed the hypothesis of self-driving cars to improve the convenience and protection of automobile users(Hamian et al.; 2021). Scene interpretation is one of the criteria for this intelligent system that can help in making the correct choice. Because it is important to analyse the pictures received from those in the cameras, so that situation can be comprehensively analysed. Semantic segmentation could be utilized as an ideal approach to tackle these problems.

The application of deep learning methods like Convolutional Neural Network(CNN) and its derived methods have shown good results in field of computer vision field, which has motivated many researchers to work in semantic segmentation that could help in addressing many real life problems. Out of which driver-less cars and disease diagnosis in medical field have attracted most attention. Various segmentation methods have been proposed throughout the years(Grace et al.; 2021). Few approaches give superior results for each sort of area, which might be uniform, contoured, or degraded, because segmentation focuses upon region-based processing. Whereas to determine the effectiveness and accuracy of segmentation results, there are two types of algorithms that are employed in this research.

The proposed architecture in this research work is practical exploration of deep learning techniques for segmentation task. CNN(Convolutional Neural Networks) have performed well for basic pictures, but not so well for complicated ones. Other algorithms, such as U-Net or Res-Net, come into effect here to target the problem. U-Net and Res-Net convolutional network are applied on camvid dataset to explore the practical effectiveness on autonomous driving through image segmentation.

1.2 Background and Motivation

The process of deducing what is there in the world from pictures is known as perception. It's a perplexing sense that gives us the most strong environmental clues. Knowledge from the environment around us is captured by our eyes at a rate of 10 gigabits per second. This information is processed by the brain at a rate of more than 3 million bit / s (Andersen et al.; 2005). Human brain can remembers a remarkable high quality image of the world surrounding us by using this information(Solina; 1991). Researchers from fields as varied as medicine, sociology, neuro-science, architecture, computer science, as well as artificial intelligence have long struggled to create an intelligent machine that can attain the same accuracy and reliability as humans(Kendall et al.; 2015).

Computer vision is an interdisciplinary field designed to allow machines visualize. Because of the tremendous complexity and diversity in appearance one can perceive the visual environment as extremely challenging and difficult. Designing a new technique has not been possible to scale to an acceptable degree of knowledge until now(Kendall; 2019). Machine learning techniques are the most promising way for developing computers that can comprehend pictures at a human level. The field of computer vision have a significant influence on many new fields of rising technology (Yi; 2021).
Being able to create intelligent vision methods may contribute significantly to our knowledge of neuro-science behind visual intelligence. Autonomous driving is challenging robotics task that involves observation, planning, as well as execution in ever-changing settings. Because safety is important, this work must also be completed with the highest accuracy. Semantic Segmentation can recognize lane lines and traffic signals, as well as offer information about open space on the road [1].

[1] https://www.kdnuggets.com/2018/10/semantic-segmentation-wiki-applications-resources.html

1.3 Research Question

How well can application of U-NET and FPN convolution neural network improves driving scenes perception using semantic segmentation?

1.4 Research Objective

Objectives of this research are as follows:

1. Analysis of the literature review in the field of image segmentation associated with application to real world problems.

2. Development and implementation of the deep learning segmentation model on driving scene dataset.

3. Evaluate the models performance on the basis of evaluation metrics for segmentation.

Contribution: The major contribution of this research work is practical exploration of convolution method in field of image segmentation for understanding the autonomous driving perception. Convolution methods, U-NET with ResNet-50 and FPN with ResNeXt-50 is developed and implemented for driving scene images using camvid dataset. The region of interest is localized and segmented using a Convolution Neural Network that employs pre-processed image patches (driving scenes). This could be helpful in further exploration of various combination of deep learning methods that could be employed to develop better segmentation results in field of autonomous driving system. This will offer information about open space and objects on the road and make it more efficient.

The research paper is organized into following sections. Section 2 scrutinizes the state-of-the-art deep learning techniques applied for semantic segmentation. The methodology for proposed model has been narrated in Section 3. Then the next section which is section 4 consist of the design specification of the proposed model used for semantic segmentation in driving images. Section 5 and 6 contains the implementation and evaluation outcome of the proposed scheme using CamVid dataset. Section 7 briefly concludes this research paper.

2 Related Work

Various techniques for image segmentation will be described with attempts to strike a balance between precision and reliability. Two of the problems which have made designs difficult to achieve excellent accuracy include correctly identifying pixel labels in image of objects and handling objects of different sizes. In past decade, deep neural networks techniques has shown potential in image processing as well as in computer vision domain such as object detection, motion detection, vegetation recognition and image segmentation. Evidently, using deep learning based semantic segmentation has improved the accuracy of identifying the items in a picture as well as their locations(Hamian et al.; 2021). The literature review is divided into two parts such as semantic segmentation and supervised segmentation methods that are applied.

2.1 Semantic Segmentation

Semantic pixel-by-pixel segmentation necessarily requires determining the semantic category of each pixel. It is a widely discussed topic, fueled by difficult data sets (Yang et al.; 2020). Prior to deep learning, the most effective methods relied on hand-crafted features that categorized pixels separately. An image piece was typically fed as input to classifier, such as random forests and boosting algorithm, to anticipate the target class of the center pixel (Brostow et al.; 2008). There have been research on characteristics based on appearance or motion and presence. Noisy prediction of each pixel which are obtained from classifiers and by using CRF (Conditional Random Field) pair-wise to enhance the accuracy (Ladickỳ et al.; 2012).

More subsequent techniques have attempted for predicting pixel labels in a patch rather than just the central pixel in order to attain greater accuracy. This improves random forest prediction performance but decreases thin structure performance (Muller and Behnke; 2014). Dense depth maps have also been classified using Random Forests (Zhang et al.; 2010). Additional viewpoints suggest for the adoption of a blending of well-known crafted features along with spatial feature-pixels to improve accuracy (Tighe and Lazebnik; 2013). After the publication of the NYU dataset, which comprises labelled RGB-D data acquired from a Kinet sensor (Ren et al.; 2012). (Muller and Behnke; 2014) show that adding depth to model increases segmentation accuracy enormously. All of these techniques, however, categorise RGB-D pictures using hand-crafted features.

Because of their effectiveness in object classification, researchers have recently begun to utilize deep convolutional neural network (DCNN) feature learning capacity for systematic prediction tasks like segmentation. Object categorization networks were also used to segment images, most notably by duplicating th the last few layer features in network to suit picture sizes (Hariharan et al.; 2014). Jiang et al. (2021) used recurrent neural networks to combine numerous low resolution estimations to produce input picture resolution predictions. These approaches are already superior than hand-engineered features, however they have limited capacity to define limits.

Semantic segmentation provide a in-depth information about a image, as compared to object classification and detection. This knowledge is critical in a variety of fields, including autonomous driving, robots, picture search engines, and so on. A plethora of semantic segmentation techniques have been recently developed (Garcia-Garcia et al.; 2017). FCN , for instance, conducted semantic segmentation via pixel-wise prediction initially, yielding substantially good performance in natural scene, in end-to-end structure(Long et al.; 2015). By incorporating more skip architecture with max-pooling module, SegNet makes the technique more efficient over FCN (Badrinarayanan et al.; 2015). Dilated convolution with pyramid pooling were utilized by PSP-Net to modify SegNet model (Xu et al.; 2021).

U-Net, which comprises of contracting and proportionally growing sub-networks that produce a U-shaped design, was suggested as in 2015 ISBI challenge. This model was created to tackle the problem of biomedical picture segmentation. It is extensively utilized in many aspects of semantic segmentation because it only takes a minimal amount of training data to obtain acceptable segmentation results (Cai et al.; 2020). All of the following semantic segmentation methods heavily rely on convolution procedures to retrieve semantic (global) as well as appearance (local) features from pictures. Nevertheless, the semantic as well as appearance data extracted via shallow convolution layers are generally restricted when segmentation task is performed on complicated images.

In general, they choose to strengthen the network layers such that the semantic segmentation model can effectively achieve the semantic and visual information of the pictures, therefore increasing segmentation model performance. But, if the network continues to deepen forever, both the processing power and optimizer will face a significant difficulty. As a result, we should address the cause of the problem through optimizing the convolutional techniques.

2.2 Supervised Segmentation Methods

In the realm of computer vision, deep learning is now a benchmark. Neural networks is particularly compelling at comprehending high-dimensional inputs, like as graphics and videos.

Yi (2021) proposes a semantic segmentation algorithm that is based on RNN(Recurrent Neural Network). By the application of recurrent network, the process of image segmentation could be smooth and enrich the details present in image or video. The model was tested on two dataset HELEN face dataset and PASCAL person parts. Research on HELEN face and PASCAL- human body data sets demonstrate that employing a recurrent neural network to do semantic segmentation improves the efficiency of fine tuned image segmentation. The model was evaluated on the basis of mean intersection of union (mIOU), where model score achieved on both dataset were 47.84% and 56.43% respectively.

Kendall et al. (2015) proposes a segNet algorithm which was based on deep convolutional network, so that it could aid architectural field for better understanding of the scene. The experiments shows that model predicts pixel-wise label of classes and Monte Carlo sampling technique was used at the time of testing to evaluate the end result. By modelling uncertainty their was slight change in efficiency of proposed model, when compared to a few state-of-the-art method such as FCN and Dilation network. The model was tested on different dataset and performance was compared to the other state-of-art method for all the available dataset. Another model was proposed by (Kendall et al.; 2015) performed best on SUN scene understanding dataset. The evaluation measure used for evaluating model were accuracy and IoU(Intersection of Union).

Weeds and pests these are two most common sources of crop damage. To achieve good yields, several conventional approaches are employed to limit the growth (Grace et al.; 2021). Environmental degradation and agricultural contamination are two main drawbacks of these techniques, both of which are harmful to human health. Grace et al. (2021) proposed a deep learning algorithm based on cascaded segmentation method which differentiates weed from the other crop and provide aid in farming to tackle this problem. The cascaded network was trained on smaller network in order to get coarse-fine feature and later these feature predictions were combined with end result to obtain segmentation results of weed and pests. The evaluated results illustrated that proposed cascaded network performed well in comparison to the other state-of-art methods like U-NET, FCN-8s and DeepLabv3 over scores of IoU, F1 score and true detection rate.

The automatic detection and image segmentation of Lymph Nodes (LNs) from the cross-sectional radiology images is a crucial move in the automatic scientific evaluation of cancer patients. Despite this, due to the poor contrast of LNs with neighbouring soft tissues and also the diversity in nodal shape, it remains a challenging process. Xu et al. (2021) came up with deep dilated cascaded framework that could help in addressing these issues. The model targets two important issues which arises in LNs segmentation such as pixel loss during training which are addressed by employing loss function such as cosine-sine(CS) and to address performance issue while segmentation process, a dilated spatial pyramid pooling is employed. The evaluation results obtained Dice similarity coefficient(DSC) score as 77% compared to other state-of-art method, where segNet DSC score was 71%.

The high computational cost of using CNNs on devices for dense prediction tasks restricts their use presented a light weighted CNN framework to somewhat aid this issues. The model utilizes an asymmetric encoder-decoder architecture for semantic segmentation. First the encoder, in particular, uses a ResNet as a backbone with two novel functions, channel split and shuffle, for each residual block that drastically lowers computing costs while retaining greater segmentation accuracy. Whereas, by adding attention pyramid network to decoder framework makes whole convolutional network light. The model was evaluated and tested on cityscapes dataset, where it achieved 71.6% mIoU overall and about 87.1% of category mIoU.

The base-line architecture of deep learning method for semantic segmentation is either based on spatial pooling module or a structured encoder-decoder system. While former method used filters to capture the features and with deep learning methods employ spatial pooling to extract features from image.Chen et al. (2018) proposed an improved version of DeepLabv3 as DeepLabv3+ which was added to decoder module to obtain segmentation of object with boundaries. Moreover, an Xception model was employed in order to enhance the model accuracy and speed. The model was tested and evaluated on PASCAL-VOC 2012 and Cityscapes dataset, where proposed model obtained accuracy of 89% and 81% respectively.

Long et al. (2015) research on fully convolution neural network (FCN) illustrates that FCN network if trained as end-to end approach, pixel-to-pixel could itself beat other state-of-art algorithms for semantic segmentation. Long et al. (2015) define and describe the space of FCN, explain how they're used to solve spatially dense segmentation problem. By taking the support of other state-of-networks like VGG-19,GoogLENET and Alex-net adapting their classification network into FCNN and fine tune the segmentation task by learning contextual knowledge via transfer learning. The model was tested and evaluated on PASCAL-VOC dataset and there was of about 20% change in performance in contrast to other state-of-art algorithms. Overall, model achieved a mIoU score of 62.2%.

Object categorization from pictures is one of several applications where deep learning techniques have proved effective. Panboonyuen et al. (2017) proposed a DCED architecture for segmenting road items from aerial pictures that has been enhanced. The model is based on few modification from actual deep learning architecture as instead of using ReLU(Rectified Linear Unit) for classification, ELU(Exponential Linear Unit) was used. Thus training dataset for model were increased by 8 and landscape metrics was used to enhance the accuracy of the model for occluding false detection of road. The metrics used for evaluating the model were precision and F1 score, were proposed segnet model outperforms other state-of-art models by obtaining accuracy of 85.4%.

Object tracking is one of modern research subject that has piqued the interest of many academicians owing to its wide-ranging applicability. In recent years, researchers have shifted from traditional crafted procedures to deep learning algorithms in order to capitalize on the reputation and achievement of deep learning based techniques in domains like computer vision, object identification, and speech recognition. Beikmohammadi et al. (2019) proposed a algorithm which could perform efficiently and provide higher accuracy via application of transfer learning in area of object recognition. The softmax classifier is being used to categorize the activities after a pre-trained deep neural network extracts information from the selected dataset. The model was tested on KTH and UCF dataset which is based on sports activities. The evaluation results showed accuracy of around 96% in segmentation task on KTH and UCF dataset.

Although semantic segmentation is extensively utilized in the creation of semantic maps, a robust method that has a drawback of poor real-time efficiency in automated applications, particularly in parking lots. Ren and Liu (2020) proposed a method that is based on PFPN (Panopatic Feature Pyramid Network) so that time taken while performing semantic segmentation could be decreased. The model is designed as to decrease the execution time for semantic segmentation, the baseline PFPN segmentation branch mechanism is cut and now the object target feature layer produced from the segmentation block is reduced. The model was evaluated on custom made dataset and results observed indicated that times reduction in performing segmentation task was reduced by 22.1%.

Both academics and business face considerable challenges in automating the digitization of paper maps. The process of map digitization is heavily reliant on manual image analysis, which is inefficient. Badrinarayanan et al. (2015) proposed a method that could aid in digitisation of the paper maps, a U-shaped convolutional neural network architecture was used for semantic segmentation. The model was tested and evaluated on map data obtained from Shibuya district of Tokyo and Jaccard similarity coefficient was used as evaluation measure of which its score on map dataset was 93.1%.

In image segmentation, multi-scale recognition is an efficient method to tackle scale variation of factors and elements. He et al. (2019) proposed multi-scale deep learning model that could effectively extract contextual features of semantic labels at pixel-level. The model is comprised of dynamically arranged convolution layers in parallel manner and every layer extract features using contextual filters for scalable semantic segmentation. The model was tested and evaluated on PASCAL-VOC 2012 dataset and the metric used for evaluation was mIoU(mean Intersection over Union), the model mIoU score was 84.4%. As compared to other state-of-art algorithms the performance was almost similar.

3 Methodology

3.1 Introduction

The techniques utilized in data mining research include Knowledge Discovery in Databases and Cross-industry standard process for data mining (CRISP-DM). KDD is a good fit for this study, since model implementation and deployment of the business logic aren't relevant. For data mining, KDD methodology is widely applied and used in various field for image segmentation and image classification such as for autonomous driving, facial segmentation, precision agriculture and geo-sensing. Various steps are comprised in KDD and each step has different functionality, such as selection of data, data preparation, preprocessing and transformation, data modeling and the last step is evaluation.

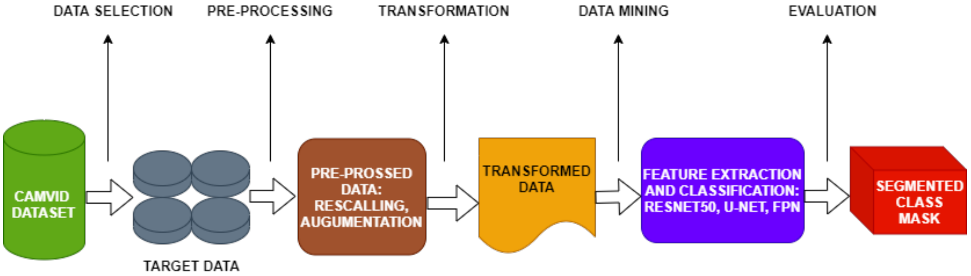


Figure 1: Methodology

3.2 Data Selection

Data selected for this research work is CamVid (Cambridge-driving Labeled Video Database) dataset. CamVid dataset is publicly available on kaggle. While, kaggle is open source platform where various dataset is available for research and learning purposes. CamVid dataset is one of the dataset which is used for research on self-driving vehicles. It contains frame by frame of road and driving scenes which was shot in 5 video segments with a 960x720 resolution camera installed on car illustrating the drive point-of view. A total of 701 frames were collected from those video sequences out of which 367 for training, 100 for validation and 234 for testing. These images are 360x480 pixels in size and includes 11 semantic class. While there are 32 classes present in the original dataset, classes have been used such car, tree, pole, cyclist, road and sky. Various scenarios from dataset for testing to get evaluation metrics of comparing models in order to execute a fair evaluation. [2]

3.3 Pre-processing and Data Transformation

Data preparation for autonomous driving semantic segmentation is done in two steps as pre-processing and transformation after acquiring the raw data file of camVid dataset from the source which is in .MXF format. After extraction the source file can be used for pre-processing and data transformation.

In the pre-processing stage, dataset selected is checked for class imbalance, resizing, normalisation, quality of data and dimensions of the images. For this research, camVid dataset consist of driving images is visualised.

While performing segmentation task, data transformation is an important step. Selected data contains less images, therefore data augmentation was performed in order to prepare data for the training purpose. Images present in data set are of size 360x480 pixel. Whereas other function such as replication of images and masking are necessary for performing segmentation task. Albumentation is used for augmentation of images, for various aspects such as Rotation, shift, magnification, flip, blur, padding, adjusting brightness and contrast. Image augmentation isn't just about data pre-processing, that includes re-scaling and normalisation of images, but it does entail some image modifications that may then be utilized for data modeling.

Different aspects in the image augmentation procedure which are necessary for training data for the deep learning techniques are performed in this research study. In initial phase batch size is defined for whole dataset. One of the key hyper-parameters for tuning

[2] kaggle datasets download -d carlolepelaars/camvid

deep learning models includes batch size. Once batch size is finalised, which is 8 in this case. Since batch size determination impacts the training of the model, if it is less or more than recommended then it results in inferior accuracy of the model. The rotation of an image is essential because there's a chance that certain pixels might be out of range, and thus the rotation procedure will retrieve all of them. Another crucial element is image flip. Images are flipped vertically in this example. Vertical flip is indeed the vertical rotation of a picture. After all of these processes the transformed data is obtained. These changes are performed on training data, which is then supplied to the U-NET and FPN segmentation model for model training.

3.4 Data Modeling

3.4.1 U-NET

Complex object characteristics, sizes, and vector object boundaries are evident in driving images. To accomplish precise segmentation, skip networks integrate high-level features from dense network layers. U-Net, which employs the skip architecture, produces excellent object area recognition outcomes. Contraction, bottleneck as well as Expansion are the three parts of design. A structure of expansion is made up of several parts of contraction. Every block has a convolution level addition input, accompanied via a maximum pooling. Each block sends data to multiple layers of CNN accompanied with a sample level, and even the lowest level mediates between the contraction layer and a expansion layer.

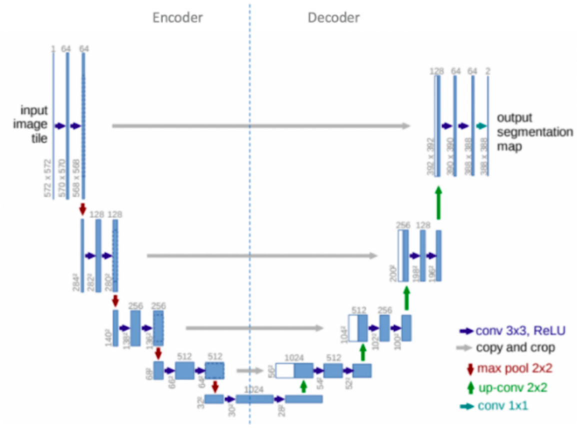


Figure 2: U-Net Work Flow Diagram
3

In the architectural diagram(Figure 1), the encoder is a first half . A pre-trained classification model like ResNet-50 is used, wherein convolution blocks and then by max-pool down-sampling it would encode its input which are driving scenes in image into feature vectors at several levels. The architecture's part 2 is the decoder. The objective is to achieve a dense classification from semantically projecting the encoder's classified features onto the pixel level. Up-sampling as well as concatenation precede ordinary convolution processes in the decoder.

3.4.2 The Feature Pyramid Network(FPN)

Object segmentation with various sizes is difficult, especially for tiny objects. Hence, a feature pyramid network is utilized to achieve driving scene segmentation for this issue. FPN is made up of top-down but also bottom-up channels. The usual deep learning network in feature extraction is the bottom-up approach. The FPN is a function eliminator that is made for pyramid model, that is built with precision and accuracy. To separate the objects within driving scenes, it eliminates the scanner function like Faster R-CNN, and creates many surface map layers with greater attribute information than the conventional recognition system. The architecture comprises of a top-down and bottom-up structure, with the bottom-up approach being a standard convolutional network for retrieving scene information and also the picture quality decreases as proceeds up. As a result, when the top layers are employed for identification of objects in driving images segmentation, especially when the objects are small in size, the outcomes are poor. FPN-ResneXt-50 model is developed and implemented on camVid driving scene dataset to acknowledge the autonomous drivers perception view through segmented mask of obtained from the images.

3.4.3 Residual Networks (ResNet50 and ResNeXt-50)

The ResNet framework was the first to incorporate the skip connection concept. Convolution, ReLU and batch normalization are generally the three stages. Where ReLu activation function act as a catalyst for making model fast and accurate. The benefits of integrating the first input x further with non-linear function F(x) assist initial layers in obtaining permission from other layers to access the differential information. To put it another way, omitting the F(x) functions allows previous layers to get a richer differential signal. As a result, this type of connectivity is chosen since it facilitates the deeper training process for segmenting various class objects in driving images. ResNeXt-50 is an improved version of ResNet-50 that proposes aggregated transformations as its main feature. The basic ResNet's three-layer convolutional section is replaced by a parallel stacking block with the same architecture. ResNeXt-50 pre-trained model is used with FPN for image segmentation on camVid dataset.

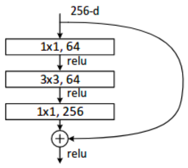


Figure 3: ResNet-50 design flow

4

3.5 Model Evaluation

The evaluation of the model for segmentation task will be performed using two evaluation metrics, dice loss and intersection over union(IoU).

3.5.1 Dice Loss

One of the most crucial aspects of every deep learning model is selecting the loss function. The Dice similarity coefficient, that is a measure of how closely two forms overlap, is a far more commonly used loss function for segmentation tasks. Even if it seems logical, the Dice Coefficient isn't the ideal choice for training. Because the dice coefficient has two possible values like 0 and 1. We won't be able to back propagate via the model's outputs, which are probabilities indicating whether or not each pixel represents a object. Zou et al. (2004). For training the network dice loss is utilize as the objective function. The loss is calculated by averaging the loss across all pixels in a mini-batch that have a valid label. When the percentage of pixels in each class with in training set varies greatly (for example, road, sky, and building pixels), it is necessary to scale the loss appropriately dependent on the true class. Thus dice loss is 1- value of dice coefficient.[5]

$$\mathcal{L}_{Dice}(p,q) = 1 - \frac{2 \times \sum_{i,j} p_{ij} q_{ij} + \epsilon}{\left(\sum_{i,j} p_{ij}^2\right) + \left(\sum_{i,j} q_{ij}^2\right) + \epsilon}$$

Figure 4: Dice Loss

3.5.2 Intersection Over Union(IoU)

IoU is a evaluation technique that is utilised to determine the efficiency of an object recognition for a given dataset. The formula of IoU is as: IoU = true positive / (true positive + false positive + false negative). As illustrated in the figure 5, IoU is the portion of overlap joining the predicted class segmentation along with the ground truth which is divided by the total portion of union of predicted class segmentation along with the ground truth of class. The efficiency of measure has a range from 0–1 (0–100%), with 0 depicting no overlap and 1 representing complete overlap. The accuracy of the model is determined from the value of IoU received after training and testing the model [6].

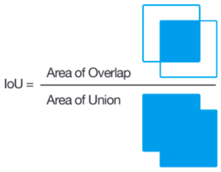


Figure 5: Intersection over Union

3.6 Segmented Class Mask

The results obtained from model implementation is explained in section 6. The model accuracy and loss based on evaluation metrics is also described.

[5] https://towardsdatascience.com/how-accurate-is-image-segmentation-dd448f896388

[6] https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2

4 Design Specification

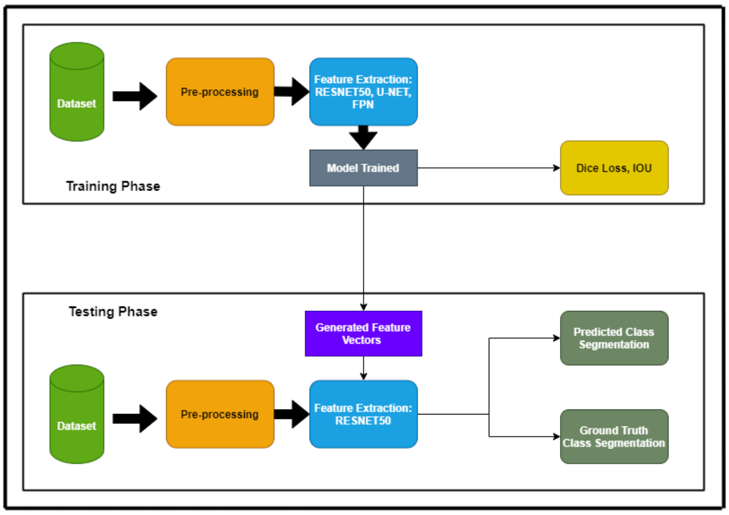

Dataset
Pre-processing
Feature Extraction: RESNET50, U-NET, FPN
Model Trained
Dice Loss, IOU
Training Phase
Testing Phase
Generated Feature Vectors
Predicted Class Segmentation
Dataset
Pre-processing
Feature Extraction: RESNET50
Ground Truth Class Segmentation

Figure 6: Design Flow Diagram

This section explains the structure and flow of the work. The most of these procedures were already covered in the Methodology part.In this part, the model's design parameters will be detailed. The implementation, assessment, then conclusion will be covered in the following sections.

Data was first imported, then pre-processed and transformed so that it could be utilized for training phase. Transformed dataset is split into three parts train, test and validation. In both the training and testing phases, the transformed images are fed into the ResNet-50 and ResNeXt-50 models, where features from the images are retrieved and a mask for each image is created. The feature maps of segmented mask pixels are obtained after training and testing phase of the model to present the forecasting of the segmented mask for object classes in form of visualised data as ground truth mask and predicted mask from the model. The ground truth and forecasted mask is presented as evaluated result for the model.

5 Implementation

In this project the computational configuration used are 1.6 GHz Dual-Core Intel Core i5 processor, memory:8 GB 1600 MHz DDR3 and Intel HD Graphics 6000 1536 MB. PyTorch segmentation models are used to build the deep learning model Subramanian (2018). This section provides insight into the use of data mining to segment driving images in detail. While training the network, deep learning models take longer to process the pictures. The experiment is conducted on a Google Co-lab PRO cloud machine with a 100 GB hard drive, 16 GB RAM as well as 72 GB run-time GPU. Pre-trained ResNet-50 and ResNeXt-50 encoders is used as backbone of the model architectures for performing image segmentation on driving scene images. Since model take longer time to run on image dataset, proper software and hardware requirement should be analysed. The use of a GPU in real time speeds up the execution of convolutional neural network. Py-torch segmentation module libraries Keras and Tensor-Flow are utilized to implement the proposed model.

5.1 Model Implemented

After a thorough examination of prior works done and suggested. Neural network methods proved to be the most efficient in field of image processing and object recognition. As a result, U-Net, FPN, as well as ResNet50 were selected to be trained on Camvid data set and then used to segment objects from driving images provided in dataset. Together with the segmented objects, the analysis of prediction is shown. The model is built on ResNet50-U-Net model that is implemented on camvid driving scene dataset. The model is based on CNN in which input of driving scene image is fed to pre-trained ResNet-50 model for extraction of the features from the training dataset and further model undergoes training.

Before training the model all important necessary libraries are installed That are required in order to train the algorithm for segmentation. Installation of libraries were done using !pip command followed by library name. Structure of library is based on py-torch module, for segmentation task network backbone is selected as ResNet-50 and its pre-trained weights are used for this project. Number of the class is defined in order to train the model. Activation function used in this project are sigmoid for classes less than two and for training model on more than two classes softmax activation function is used. Adam optimiser is also set for smooth training of model. After training the model generated feature vectors are stored for specified category. Then in testing phase, model is tested with those feature vector on test dataset of driving images and produces the segmented result for the particular class in form of masked region which omits the selected class and mask the background. The results were analysed based on ground truth masked region and predicted class masked region.

5.1.1 Setting Parameters

The research has the total number of epochs empirically set at 20 and batch size for training is set at 8 and for validation is 1. Training data was augmented prior to the model training to make it more compatible with sample data, that will improve the model prediction performance. Adam optimiser is used as optimiser, Since for training the model Adam optimiser fuses the finest attributes of the Ada-Grad and RMS-Prop algorithms and advance for an optimization solution in case of noisy situations with sparse gradients. Setting of learning rate rate is also an important aspect that will training the deep learning model. Furthermore, a network with strong generalization qualities should be resilient, meaning that tiny difference in the channel's parameters should not result in significant performance difference. Therefore, learning rate is set as lr = 1e-5. [7]

5.1.2 U-NET

For building a deep learning method for segmentation task, some libraries needs to install such as segmentation models py-torch module which consist all required libraries such tensor-flow, sci-kit learn, numpy, matplot, data loader to build an efficient model. Once all library requirements are satisfied, pre-processing and transformation of data is done. From training data, input image and respective segmented mask is fed to the network for training the model and implementing it. For segmentation task using ResNet-50 as backbone, U-Net model is implemented for performing segmentation on camvid dataset.

[7] https://www.jeremyjordan.me/nn-learning-rate/

U-Net is a FCNN that can segment images semantically. The encoder along with decoder components are linked through skip connections.The encoder extracts features with varying spatial levels (skip connections), which the decoder uses to create an efficient segmentation mask. For fusing decoder sections with skip connections, concatenation is used. The encoder for the segmentation model is ResNet-50, which extracts features of various spatial resolutions from driving images set as input. After selection of encoder, encoder range is set for this project number of steps is step to 40 and encoder range is set as (0,40). Since each step generates characteristics that are twice as small in spatial dimensions as the one before it. After setting the encoder range encoder weight is set, which in this study is set as "image-Net". Then number of classes on which encoder will extract features is defined and activation function is set accordingly, sigmoid for classes less than two and for training model on more than two classes softmax activation function is used for performing multi-class segmentation.

Once model is trained, a trained model .pth file is generated and saved as checkpoint and results evaluation graph is generated based on dice-loss and IoU value. Model goes under testing phase and once training phase is done. At last, segmentation results are visualised based on ground truth and predicted mask for particular class.

5.1.3 FPN

Object segmentation with various sizes is difficult, especially for tiny objects. We utilize a feature pyramid network (FPN) to achieve Driving scene segmentation for this issue. FPN is made up of top-down but also bottom-up channels. The usual deep learning network in feature extraction is the bottom-up approach. The Feature Pyramid Network (FPN) is a function eliminator that is made for pyramid model, that is built with precision and accuracy. The model is also performed using FPN design flow for segmentation task using encoder as Variant of ResNeXt-50 as backbone, which will perform feature extraction from the input images and feed the generated feature vector of object classes with segmentation mask to the decoder than model will be trained by defined the encoder range.

After setting input channel parameter which will create direction for model building and process tensor for numbers of arbitrary channels. Since we have used pre-trained weight ResNeXt-50 for segmentation model, convolution weights will be randomly initialised and decoder will use the encoder information on dataset to validate the model. The encoder for the segmentation model is ResNeXt-50, which extracts features of various spatial resolutions from driving images set as input. After selection of encoder, encoder range is set for this project number of steps is step to 40 and encoder range is set as (0,40). After setting the encoder range encoder weight is set, which in this study is set as "image-Net". Then number of classes on which encoder will extract features is defined and activation function is set accordingly, sigmoid for classes less than two and for training model on more than two classes softmax activation function is used for performing multi-class segmentation. After model is trained, testing of the model is done on test dataset and model is validated by comparing the feature vector obtained from the trained model on the basis of ground truth segmented mask and prediction mask for specified class used for segmentation.

Once model is trained, a trained model.pth file is generated and saved as checkpoint and results evaluation graph is generated based on dice-loss and IoU value. Model goes under testing phase and once training phase is done. At last, segmentation results are visualised based on ground truth and predicted mask for particular class.

6 Evaluation

Model evaluation is one of important steps to analyse, if the proposed task is performed well or not. For exploring the image segmentation task on camvid dataset, two models were applied one with U-Net architecture and other with FPN-ResNeXt-50 to achieve the segmented mask result for selected classes. Evaluation measures used for both models were dice loss and intersection over union(IoU) and results obtained after model training and testing are analysed. Then using these measures scores, ground truth mask and predicted mask of the selected class segmentation is observed. The table below summarizes the results of both model for image segmentation on driving scene image dataset(camvid).

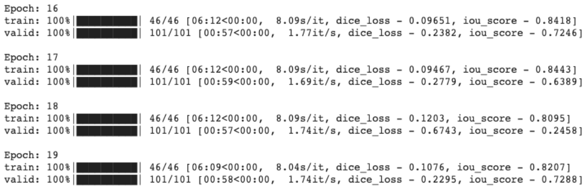
Model	**IoU Score**
U-Net-Resnet-50	82.02%
FPN-ResneXt-50	84.7%

Table 1: Results Summary

from the table above we can observe that results obtained by U-Net-50 model is 0.82 in terms of IoU score. Whereas, FPN-ResNeXt-50 model gave IoU score of 0.847. Other visualised results are explained in subsection of evaluations. Where segmentation results and graphs are explained.

6.1 U-Net-ResNet-50 model

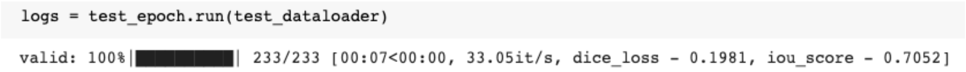
Evaluation results of the model are explained in this section. Training of the model was done at 20 epoch, where U-Net-ResNet-50 model obtained a IoU(intersection over union) score of 82.07% and 72.88% on validation of model on camvid dataset. As shown in figure 7 below.



```
Epoch: 16
train: 100%|██████████| 46/46 [06:12<00:00,  8.09s/it, dice_loss - 0.09651, iou_score - 0.8418]
valid: 100%|██████████| 101/101 [00:57<00:00,  1.77it/s, dice_loss - 0.2382, iou_score - 0.7246]

Epoch: 17
train: 100%|██████████| 46/46 [06:12<00:00,  8.09s/it, dice_loss - 0.09467, iou_score - 0.8443]
valid: 100%|██████████| 101/101 [00:59<00:00,  1.69it/s, dice_loss - 0.2779, iou_score - 0.6389]

Epoch: 18
train: 100%|██████████| 46/46 [06:12<00:00,  8.09s/it, dice_loss - 0.1203, iou_score - 0.8095]
valid: 100%|██████████| 101/101 [00:57<00:00,  1.74it/s, dice_loss - 0.6743, iou_score - 0.2458]

Epoch: 19
train: 100%|██████████| 46/46 [06:09<00:00,  8.04s/it, dice_loss - 0.1076, iou_score - 0.8207]
valid: 100%|██████████| 101/101 [00:58<00:00,  1.74it/s, dice_loss - 0.2295, iou_score - 0.7288]
```

Figure 7: Training result for U-Net-ResNet-50

Model was trained on 20 epoch, as seen in figure 7. Evaluation measures can be observed for training model, where training and validation feature loss is determined by dice-loss score which was 0.10 and 0.229 for model which shows model is a good fit, as loss of features while training the model is not so much segmented mask region for class car can be predicted well. Results obtained shows that less loss value indicate less feature pixels are lost while generating the segmented mask for training data. Thus indicates that model was trained well. As shown in figure 7 above.



```
logs = test_epoch.run(test_dataloader)

valid: 100%|██████████| 233/233 [00:07<00:00, 33.05it/s, dice_loss - 0.1981, iou_score - 0.7052]
```

Figure 8: Training result for U-Net-ResNet-50

After testing the model, dice loss and IoU was recorded as 0.283 and 75 displayed in figure 8. Which is a Fairly good evaluation score. But loss of 0.28 could result in loss of feature pixel while segmenting mask for class car.

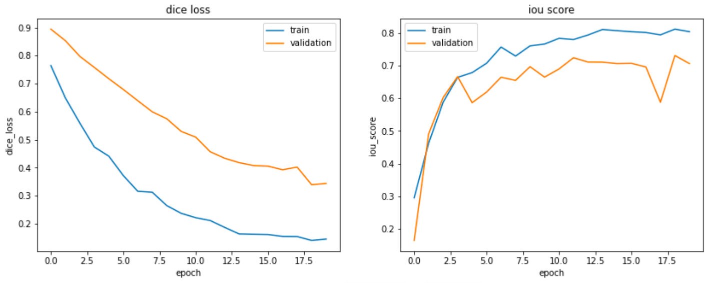


Figure 9: Test result for U-Net-ResNet-50

From figure 9, it can be observed that at the start of model training loss of feature pixels were high for training data and validation. Around 90% for validation and 78% for training data, as number of epoch increases the resultant loss decreases for training and validation. Model is run 20 epoch for these results, at 20 epoch the final value for dice loss was at 0.10 and 0.229. Whereas for IoU score in the beginning of model training was less as can be seen in graph above, but at end of the epoch cycle the IoU score for training and validation stands at 0.82 and 0.728.

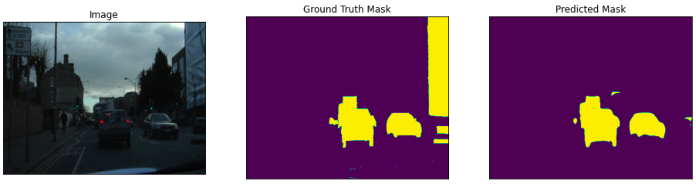


Figure 10: Segmented mask result for U-Net-ResNet-50

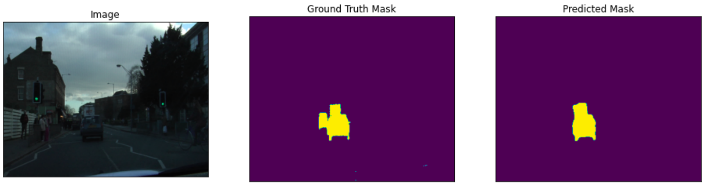


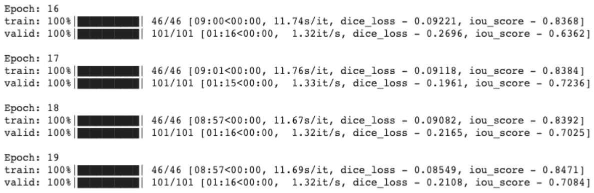
Figure 11: segmented mask result of class building for U-Net-ResNet-50

As from figure 10 and 11 above, it can be deduced that model has performed well in predicting pixels of the class and it is presented in form of ground truth mask which is the actual masked pixel area of selected class ’car’ and predicted mask represents the model output in form of segmented mask. Seeing from image, it can be said that model performed well in segmenting pixel of background(remaining pixels) and foreground(selected class pixels). As model was validated on random sample of images for 5 iterations to analyse the result some of them are presented.

. . .

6.2 FPN-ResNeXt-50

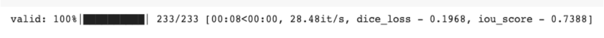
The results obtained after successful implementation of the FPN-ResNeXt-50 model on camvid dataset were observed as dice loss score for training and validation were 0.085 and 0.21 respectively which are better than U-Net-ResNeXt-50 model. Whereas, IoU score for training and validation were 0.84 and 0.70. For testing part dice loss and IoU were 0.19 and 0.73. All these results were achieved when model was trained at 20 epoch and learning rate was set at 1e-5.



```
Epoch: 16
train: 100%|██████████| 46/46 [09:00<00:00, 11.74s/it, dice_loss - 0.09221, iou_score - 0.8368]
valid: 100%|██████████| 101/101 [01:16<00:00,  1.32it/s, dice_loss - 0.2696, iou_score - 0.6362]

Epoch: 17
train: 100%|██████████| 46/46 [09:01<00:00, 11.76s/it, dice_loss - 0.09118, iou_score - 0.8384]
valid: 100%|██████████| 101/101 [01:15<00:00,  1.33it/s, dice_loss - 0.1961, iou_score - 0.7236]

Epoch: 18
train: 100%|██████████| 46/46 [08:57<00:00, 11.67s/it, dice_loss - 0.09082, iou_score - 0.8392]
valid: 100%|██████████| 101/101 [01:16<00:00,  1.32it/s, dice_loss - 0.2165, iou_score - 0.7025]

Epoch: 19
train: 100%|██████████| 46/46 [08:57<00:00, 11.69s/it, dice_loss - 0.08549, iou_score - 0.8471]
valid: 100%|██████████| 101/101 [01:16<00:00,  1.32it/s, dice_loss - 0.2108, iou_score - 0.7084]
```

Figure 12: Training result of class car for FPN-ResNeXt-50

Training of the model was done at 20 epoch, where FPN-ResNeXt-50 model obtained a IoU(intersection over union) score of 84.71% and 70.8% on validation of model on camvid dataset. Whereas loss of feature pixel while training and validation is determined by dice-loss score which was 0.085 and 0.21 for training and validation of the model, which shows model is a good fit. Results obtained shows that loss was less thus indicates that model was trained well. as shown in figure 12 below.



```
valid: 100%|██████████| 233/233 [00:08<00:00, 28.48it/s, dice_loss - 0.1968, iou_score - 0.7388]
```

Figure 13: Test result of class car for FPN-ResneXt-50

FPN-ResNeXt-50 model test results are shown in figure 13. Dice loss value as 0.19 and IoU score as 0.78. From obtained results, it can observed that 0.21 value dice loss indicate percentage pixel feature is lost while segmenting mask to the object class and value of IoU which came at 20 epoch represents that 73.8% area of the object class is detected and segmented successful. Which in turn is a better result for the model.

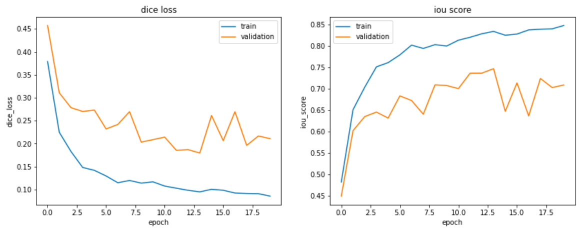


Figure 14: Training result of class car for FPN-ResneXt-50

From figure 14, it can be observed that at the start of model training loss of feature pixels was high for training data and validation. Around 45% for validation and 38% for training data, as number of epoch increases the resultant loss decreases for training and validation. Model is run 20 epoch for these results, at 20 epoch the final value for dice loss was at 0.085 and 0.21. Whereas for IoU score in the beginning of model training was less as can be seen in graph above, but at end of the epoch cycle the IoU score for training and validation stands at 0.84 and 0.708.

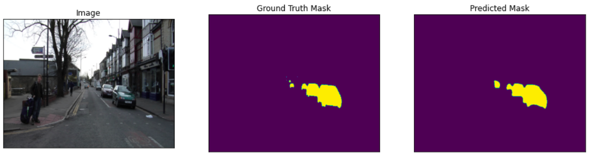


Figure 15: Segmented mask result of class car for FPN-ResNeXt-50

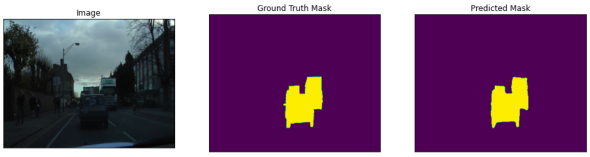


Figure 16: Segmented mask result of class car for FPN-ResneXt-50

As from figure 15 and 16 above, it can be deduced that model has performed well in predicting pixels of the class and it is presented in form of ground truth mask which is the actual masked pixel area of selected class 'car' and predicted mask represents the model output in form of segmented mask. Seeing from image, it can be said that model performed well in segmenting pixel of background(remaining pixels) and foreground(selected class pixels). As model was validated on random sample of images for 5 iterations to analyse the result few are presented here.

6.3 Discussion

Results obtained after model implementation are explained above. Both models performed well in segmentation of the driving scenes. FPN-ResNeXt-50 attained IoU value as 84% which was relatively higher than U-Net architecture which was implemented with ResNet-50 model and IoU score was 82%. Overall dice loss observed after implementation of both model were relatively low and from segmentation mask generated clearly shows that model effectively masks the target class. This work was aimed to provide a way of using convolutional neural network that could help in understanding the driving scenes through image segmentation. The implemented models are helpful in enhancing segmentation accuracy as well as achieving a decent fit.

Similar kind research was conducted by Badrinarayanan et al. (2015) in digitisation of maps for few district of Tokyo, where model used FPN architecture for image segmentation of the map and model achieved the IoU score of 93.1%. Cai et al. (2020) organised a research in biomedical field for the diagnosis of diseases and tumor where model achieved an accuracy of about 85%.

7 Conclusion and Future Work

The application of deep learning methods such as CNN and its derived methods have shown good results in field of computer vision field. Autonomous driving involves observation, planning, as well as execution in ever-changing situation, safety and accuracy are important factors. The focus of this study was to use image segmentation with implementation of deep learning in order to aid in better understanding of driving scene perception in order to help autonomous driving systems in distinguishing between ground reality and prediction. Whereas, deep learning models such U-Net, FCN and FPN are best among state-of-art method that are used for providing solutions to real world problems. For this research U-Net architecture and FPN architecture were used on camvid driving scene dataset and image segmentation was used to understand the driving scenes to predict the difference between ground reality and machine predicted result.

Both model used pre-trained encoders for segmenting mask results on camvid dataset. where U-Net with ResNet-50 encoder provided a IoU score of 0.82 and FPN with ResNeXt-50 obtained IoU score of 0.84. Thus results of both model indicate good performance in segmentation of predicted class and ground truth, which is around the accuracy of the state-of-art method used for segmentation to solve real world problems.

As model was trained on less number of epoch, loss of segmented pixels was about 20% which is large in case of real-time system. Thus model could be enhanced with application of better optimisation techniques. Further to provide better solution and improvement in driving perception application of transfer learning could be employed with state-of-art methods in future.

8 Acknowledgement

I am swamped in all humbleness and gratitude to all those who have helped me to place these ideas, well on top of the amount of simplicity and into a concrete research. I would like to express a special thanks of gratitude to my professor Mr. Noel Cosgrave who supported me during this project on Semantic Image Segmentation. Doing research work can be daunting and there were times that I almost gave up but with the assistance of my professor, I was able to put my best foot forward. My attempts at any level couldn't be satisfactorily completed without the support and guidance of my friends for their honest reviews and advice.

References

Andersen, N. A., Braithwaite, I. D., Blanke, M. and Sorensen, T. (2005). Combining a novel computer vision sensor with a cleaning robot to achieve autonomous pig house cleaning, *Proceedings of the 44th IEEE Conference on Decision and Control*, IEEE, pp. 8331–8336.

Badrinarayanan, V., Handa, A. and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling, *arXiv preprint arXiv:1505.07293* .

Beikmohammadi, A., Faez, K., Mahmoodian, M. H. and Hamian, M. H. (2019). Mixture of deep-based representation and shallow classifiers to recognize human activities, *2019*

5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), IEEE, pp. 1–6.

Brostow, G. J., Shotton, J., Fauqueur, J. and Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds, *European conference on computer vision*, Springer, pp. 44–57.

Cai, S., Tian, Y., Lui, H., Zeng, H., Wu, Y. and Chen, G. (2020). Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network, *Quantitative imaging in medicine and surgery* **10**(6): 1275.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation, *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V. and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation, *arXiv preprint arXiv:1704.06857* .

Grace, R. K. et al. (2021). Crop and weed classification using deep learning, *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **12**(7): 935–938.

Hamian, M. H., Beikmohammadi, A., Ahmadi, A. and Nasersharif, B. (2021). Semantic segmentation of autonomous driving images by the combination of deep learning and classical segmentation, *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, IEEE, pp. 1–6.

Hariharan, B., Arbeláez, P., Girshick, R. and Malik, J. (2014). Simultaneous detection and segmentation, *European conference on computer vision*, Springer, pp. 297–312.

He, J., Deng, Z. and Qiao, Y. (2019). Dynamic multi-scale filters for semantic segmentation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3562–3572.

Jiang, D., Li, G., Tan, C., Huang, L., Sun, Y. and Kong, J. (2021). Semantic segmentation for multiscale target based on object recognition using the improved faster-rcnn model, *Future Generation Computer Systems* **123**: 94–104.

Kendall, A. (2019). *Geometry and uncertain in deep learning for computer vision.*, PhD thesis, University of Cambridge, UK.

Kendall, A., Badrinarayanan, V. and Cipolla, R. (2015). Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arxiv 2015, *arXiv preprint arXiv:1511.02680* .

Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W. and Torr, P. H. (2012). Joint optimization for object class segmentation and dense stereo reconstruction, *International Journal of Computer Vision* **100**(2): 122–133.

Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Muller, A. C. and Behnke, S. (2014). Learning depth-sensitive conditional random fields for semantic segmentation of rgb-d images, *2014 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 6232–6237.

Panboonyuen, T., Vateekul, P., Jitkajornwanich, K. and Lawawirojwong, S. (2017). An enhanced deep convolutional encoder-decoder network for road segmentation on aerial imagery, *International conference on computing and information technology*, Springer, pp. 191–201.

Ren, L. and Liu, Y. (2020). Research on the application of semantic segmentation of driverless vehicles in park scene, *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*, IEEE, pp. 342–345.

Ren, X., Bo, L. and Fox, D. (2012). Rgb-(d) scene labeling: Features and algorithms, *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 2759–2766.

Solina, F. (1991). Pattern recognition and computer vision in slovenia-an overview.

Subramanian, V. (2018). *Deep Learning with PyTorch: A practical approach to building neural network models using PyTorch*, Packt Publishing Ltd.

Tighe, J. and Lazebnik, S. (2013). Superparsing, *International Journal of Computer Vision* **101**(2): 329–349.

Xu, G., Cao, H., Udupa, J. K., Tong, Y. and Torigian, D. A. (2021). Disegnet: A deep dilated convolutional encoder-decoder architecture for lymph node segmentation on pet/ct images, *Computerized Medical Imaging and Graphics* **88**: 101851.

Yang, G., Rota, P., Alameda-Pineda, X., Xu, D., Ding, M. and Ricci, E. (2020). Variational structured attention networks for dense pixel-wise prediction.

Yi, L. (2021). A progressive image semantic segmentation method using recurrent neural network, *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, IEEE, pp. 765–768.

Zhang, C., Wang, L. and Yang, R. (2010). Semantic segmentation of urban scenes using dense depth maps, *European Conference on Computer Vision*, Springer, pp. 708–721.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells III, W. M., Jolesz, F. A. and Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports, *Academic radiology* **11**(2): 178–189.