Institute for Integrated Signal Processing Systems (ISS)
RWTH Aachen University
Prof. Dr.-Ing. Gerd Ascheid

Master Thesis M107

Deep Learning for Autonomous Driving using End-to-End learning and Data Fusion

by

Deepak Ramani

Matr.-No. 314726

September, 2020

Supervisors:
Prof. Dr.-Ing. Gerd Ascheid
Shawan Taha Mohammed

I assure that this project was accomplished by me, without any foreign assistance except the official support of the chair. The used literature is fully indicated in the bibliography.

Aachen, Sep 30, 2020

(Deepak Ramani)

FIXME:

Placeholder

Please replace this sheet by the official thesis title/task document signed by the professor and you.

Contents

1 Introduction

The last decade has seen massive growth in the field of Autonomous Driving, primarily due to proliferation of graphical processing unit(GPU), and several projects like Google(Waymo) [1], Berkeley-DeepDrive [2], Apollo [3], making their datasets open-source which have made it easier for people to work on these data and achieve better performance gains.

Training a deep neural network(DNN) forms the core of making a car autonomous. By using supervised learning, one can achieve reliable results as it gives greater control at each stage of training. The data-driven approach collects data in advance and labels it appropriately. It can then be fed to the DNN using supervised learning algorithms to train the best model possible.

Ever since the discovery of Alexnet in 2012 [4], the convolutional neural network(CNN) and deep learning(DL) are preferred choices to analyse images. However, it is well known that the camera sensors are susceptible even to a slight change in weather conditions. Sensors like radar [5], LIDAR [6], ultrasonic[7], depth camera give additional depth information for obstacle detection. These features are then fused with the camera images and this process is called data fusion.

Even though there are some public data available, it is still not enough to reliably train a DNN. Then there is the cost of building an autonomous car. Fortunately, the last years have seen growth in reliable simulators which helped massively to collect data to help explore this field of research. To name a few simulators that are being actively used – LGSVL [8], Nvidia Drive [9], Carla [10], CarMaker [11]. In this thesis, the LGSVL simulator is used.

Figure 1.1: LGSVL[8] simulator active with all sensors

The LGSVL simulator allows the use of different sensors with minimal effort. The data from different sensors are published through websocket. So to capture these data, we need an interface/protocol

which can understand the sent data's message type and enable the receiving node to store them. However, the data from each sensor arrives at different rates. Hence it is necessary to collect and synchronise them in the order of their arrival before storing, so as to not lose their integrity and thereby prevent corrupting the dataset. Robotic operating system(ROS) [12] and its functionalities fulfil this purpose. It allows seamless transfer of simulator's data by subscribing to sensor nodes in the form of topics. Then the subscribing node with the help of ROS libraries, synchronises it as necessary for storage.

So, the data that resembles real-world is stored locally for later analysis and research.

1.1 Motivation

The motivation for this thesis is to use a simulator, do the required tests, and determine whether using a simulator does indeed help in perceiving the environment and accomplishing the goal of driving in the real-world.

One of the major obstacles in autonomous driving is the cost associated with integrating sensors in addition to manufacturing a vehicle. Representing the environment around the vehicle(ego vehicle) requires information from all in-car sensors. The resources demanded to make an optimal decision are also a challenge.

The high cost of associated sensors such as LIDAR[13], has put off many smaller research groups from using them in their work. Simulation allows conducting adequate tests(at a low cost) and quicker development of algorithms. Simulator provides a safer environment to test and debug these algorithms.

So with the help of a simulator it is observed how different constellation of sensors work, how different modalities interact with each other, and what impact these factors have on the overall performance of the deep neural networks.

Finally, an end-to-end system is implemented which simulates real-world behaviour and gives results which can then be applied to future research to make it more robust.

1.2 Goal

The desired goals of this thesis are listed below:

1. Building a basic autonomous driving framework which comprises of the following three components:
    - ROS - use ROS2 to synchronise the data received from the simulator and its plugin through a rosbridge, use functionalities such as slop and cache, to sort the data according to their received time in order not to scramble the information. During the evaluation, use the same functionalities to send command controls back to the simulator.
    - Rosbridge - use a bridge transport protocol that connects the ros to the simulator.
    - Docker - set up a work environment that is independent of hardware or operating system which allows easy running of the commands for data collection and evaluation.
2. Implement a system that can efficiently collect and label data.
3. Implement an end-to-end neural network architecture which applies state of the art deep learning techniques to learn driving by predicting the steering and other control commands from image pixels.
4. Implement and analyse different constellation of sensors with different data fusion techniques.

1.3 Related Work

In 2012, Alexnet [4] used CNNs to do object classification which, then in Computer Vision became the dominated approach for classification. Both Chen *et al.* [14] and Bojarski *et al.* [15] extended [4]'s approach of using CNN and showed that in addition to classification, CNN can extract features from images. Then they went on to demonstrate through an end-to-end network(which self-optimises itself based on its inputs), that steering angles can be predicted to keep the car in the lane of a road.

In a different field, but using CNN, Sergey Levine *et al.* [16] in 2016 corroborated that it was indeed possible to extract features with CNN and predict motor control actions in *object picking robots*.

Then, Xu *et al.* [17] in the same year with CNN-LSTM architecture showed that using the previous ego-motion events helped predict future ego-motion events. Using CNNs in an end-to-end architecture raised some questions on how it reached its decisions. So in 2017, both [18], [19] did visual analysis after the CNN layers to better understand the module's functionality. Vehicle control is more than just steering control. For smoother control, acceleration and braking are necessary besides steering. Both acceleration and deacceleration are dependent on the user's driving style, lane speed limit and traffic etc. Yand *et al.* [20] used CNN-LSTM architecture and provided the LSTM with feedback speed to determine the velocity of the ego vehicle.

Besides vehicle control, perceiving the environment is necessary for collision avoidance. The RGB colour camera sensors don't provide the depth information which is critical for collision avoidance. Hence, it is essential to fuse other sensors with diverse modalities with RGB to predict an optimal output. Liu *et al.* [21] provided rules in fusing data. They said that it was essential to pick out only vital information and discard other noisy data. They also described the techniques involved in data fusion – early/late fusion, hybrid fusion, model ensemble and joint training. Park *et al.* [22] gave us methods to enhance the features by using feature amplification or multiplicative fusion. Zhou *et al.* [23] detailed how fusing data into CNN affects the overall performance.

Even though the fused dataset gives a performance boost, it performs worse compared to individual modality. The combined fused model overfits more than its counterparts. The fundamental drawback of *gradient descent* in backpropagation causes the networks to overfit. This paper [24] introduced a technique called *gradient blending* to counteract this problem.

Xiao *et al.*[25] applied all the fusion techniques mentioned above with an imitation based end-to-end network[26]. They concluded that RGB images with depth information(obtained through a different modality) could indeed result in better performing end-to-end network model.

1.4 Contribution

For supervised learning task, it is essential to collect data and label them. LGSVL, ROS provide an excellent platform to carry out this work.

So with their help, it was possible to acheive

1. For first time implementation of a basic autonomous driving framework based on ROS and LGSVL.
2. An implementation of an end-to-end training neural network with LGSVL for the first time.
3. Fusing of data with different sensor in different constellations with LGSVL.

2 Fundamentals

2.1 Machine learning: What and why?

Machine learning is all about learning from data and gaining knowledge from it. Machine learning was initially thought of as automating redundant human tasks and later developed into something that allowed solving complex mathematical problems. It was seen just an addition to humans than extension of them. Machine learning these days are required to perform tasks that are quite obvious and natural to humans such as recognising faces in images or perceiving the road environment around the vehicle and making decisions instinctively.

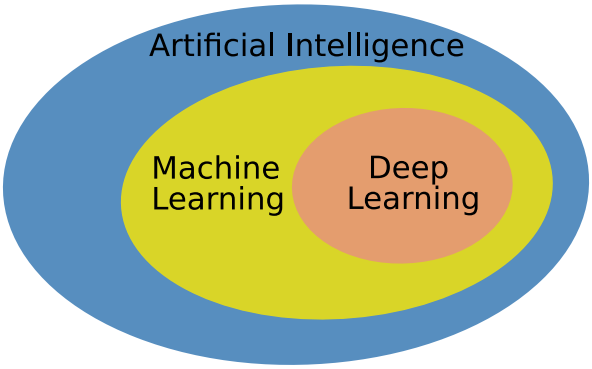


Figure 2.1: Schema of AI, ML and DL

All these attributes require to extend the field of machine learning.The figure 2.1 shows how artificial intelligence(AI) which was just a robot with simple if-else conditions, paved way for a subset in Machine learning(ML) and ML in turn getting narrower focus to result in another subset in Deep learning(DL).

So, in this chapter, a brief overview is given on the concepts that are implemented in the later chapters.

2.1.1 Learning algorithms

Machine learning provides a means to tackle tasks that are complex to solve through fixed programmes and designed by human beings [27]. A learning algorithm is an algorithm which gains the ability to learn from data. A ML algorithm is one that gains the ability to learn from an experience E with respect to some class of tasks T and performance measure P [28]. With experience, the algorithm can improve its performance.

Tasks T

The two major tasks in ML are *classification* and *regression*.

In classification related tasks, the system identifies which of k categories an input belongs to.

A function $f : \mathbb{R}^n \rightarrow \{1, \ldots, k\}$ is used by the learning algorithm to solve this task. When $y = f(x)$, the model assigns an input described by vector $\mathbf{x}$ to a category identified by numeric code y. There are other variants of the classification task, for example, where f outputs a probability distribution over classes [29]. Alexnet [4] is one of the examples of classification task that used it to do object recognition.

Regression predicts continuous value output and at any given time for an appropriate input to the neural network, regression will output a value corresponding to it.

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ predicts a numerical value for some input. Predicting the steering control value is a prime example for a regression task.

There are of course other tasks but only classification and regression are used in this thesis. Hence the narrow focus.

Performance measure P

To evaluate the performance of a ML algorithm, it is a must to design quantitative measure of its performance. Usually this performance measure P is specific to a task T. There are two distinct types of measurements – accuracy and error rate.

If the goal is to learn a mapping from inputs x to outputs y, where $y \in \{1, \ldots, C\}$, with C being the number of classes. If $C = 2$, this is called binary classification (in which case we often assume $y \in \{0, 1\}$); if $C > 2$, this is called multi class classification. If the class labels are not mutually exclusive (e.g., somebody may be classified as tall and strong), we call it multi-label classification [30].

Accuracy is just the proportion of examples for which the model produces the correct output. So in the case of binary classification, if the function f predicts a probability densities $\hat{y} \in \{0.3, 0.7\}$, for a ground truth y of value 1, then P is 70% accurate or the error rate is 30%.

It is essential that the model is evaluated with a data that it has not seen before. This data *testing set*, gives a good judgement on the performance of the trained model.

Experience E

The ML algorithms can be classified into *supervised*, *unsupervised* and *reinforcement* learning based on the kind of experience they are allowed to have. A learning algorithm is allowed to gain experience by going through the *dataset*. A dataset is collection of all the examples for a given task. For example, to classify which category a shown image belongs to has collection of images as dataset [31]. Sometimes datasets are also called as *data points*.

The focus will be on supervised learning in our case. A random vector $\mathbf{x}$ explicitly attempts to learn the probability distribution $p(\mathbf{x})$ and predicts $\mathbf{y}$ from $\mathbf{x}$, usually estimating $p(\mathbf{y} \mid \mathbf{x})$. The CIFAR dataset [31], for example, contain images as features which inturn have *targets* or *labels* associated with it. Here supervised learning(SL), the target functionality(labels) is known. So it uses the images and predicts the probability distribution to classify the images in the corresponding label.

2.2 Deep Learning

Deep learning is a subset of machine learning. It takes all the algorithms, concepts from machine learning, and narrows the focus to enable a model to learn from data such that tasks involve less human involvement, huge amount of data, and parameters.

2.2.1 Simple neural network

Linear regression is one of the common SL algorithms. It solves the regression problem. For example, if there is vector $\mathbf{x} \in \mathbb{R}^n$ as input and predict a scalar value $y \in \mathbb{R}$ as its output, then in linear

regression, output is a linear function of the input. We can define it as

$$\hat{y} = \mathbf{w}^T \mathbf{x} \tag{2.1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is a vector of parameters.

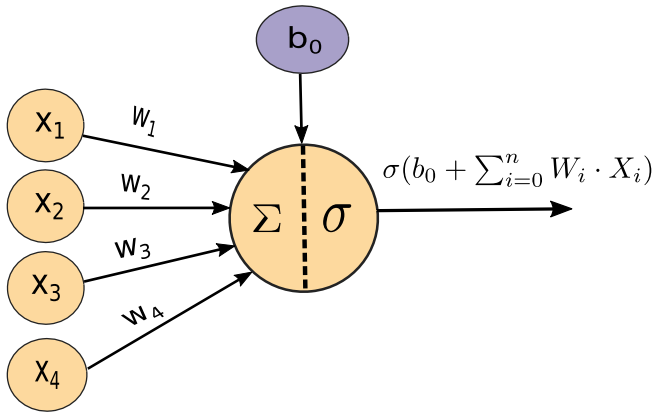


Figure 2.2: A simple neutral network

$\mathbf{w}$ is usually referred to as a set of weights that determine how each feature affects the prediction. A $\mathbf{w}_i$ is simply multiplied with a feature x_i to predict $\hat{y}$. By manipulating the $\mathbf{w}_i$ value, the corresponding feature x_i has an effect on the prediction $\hat{y}$.

A learning algorithm, in this case linear regression, is implemented as a perceptron. It is a single-layer neural network as first suggested by Rosenblatt in 1958. They generally consist of four main parts – input nodes x_i, weights w_i, bias b_0(if necessary), net sum Σ and an activation function σ. This is shown in the figure 2.2.

2.2.2 Activation function

The common activation functions used are Rectified Linear unit(ReLu), Sigmoid, tanh and softmax function. For each type of activation, σ then decides if the input received is relevant or not relevant. To convert linear inputs to non-linear, all that has to be done is to use a non-linear activation function. The figure 2.3, shows the characteristics of some of the activation functions.

For classification tasks, usually the last layer of the networks is equipped with softmax activation layer. This function normalises the output to a probability distribution over predicted output classes.

2.2.3 Multilayer feedforward networks

Deep feedforward networks or multilayer perceptrons are the quintessential deep learning models. Its goal is to approximate function f^*. In the below figure 2.4, information flows from inputs $\mathbf{x}$ to output y using a mapping function $\mathbf{y} = f(\mathbf{x}; \theta)$ where θ are the parameters values which the MLP learns for optimal approximation.

They are called feedforward as there are no feedback connections in which outputs of the model are fed back into itself. Feedforward networks with feedbacks are called *recurrent neutral networks*.

Feedforward networks form the core for many commercial applications. For example, the convolutional neural networks used for object detection are a special kind of feedforward networks.

More the hidden layers, more the depth of the feedforward networks. The width is given by the dimensionality of the hidden layer.

2.2.4 Loss function

As mentioned before, a mapping function f noisily approximates the input x to output y. So, the noise or the deviation from the true value(ground truth) must be kept at minimum. The function that

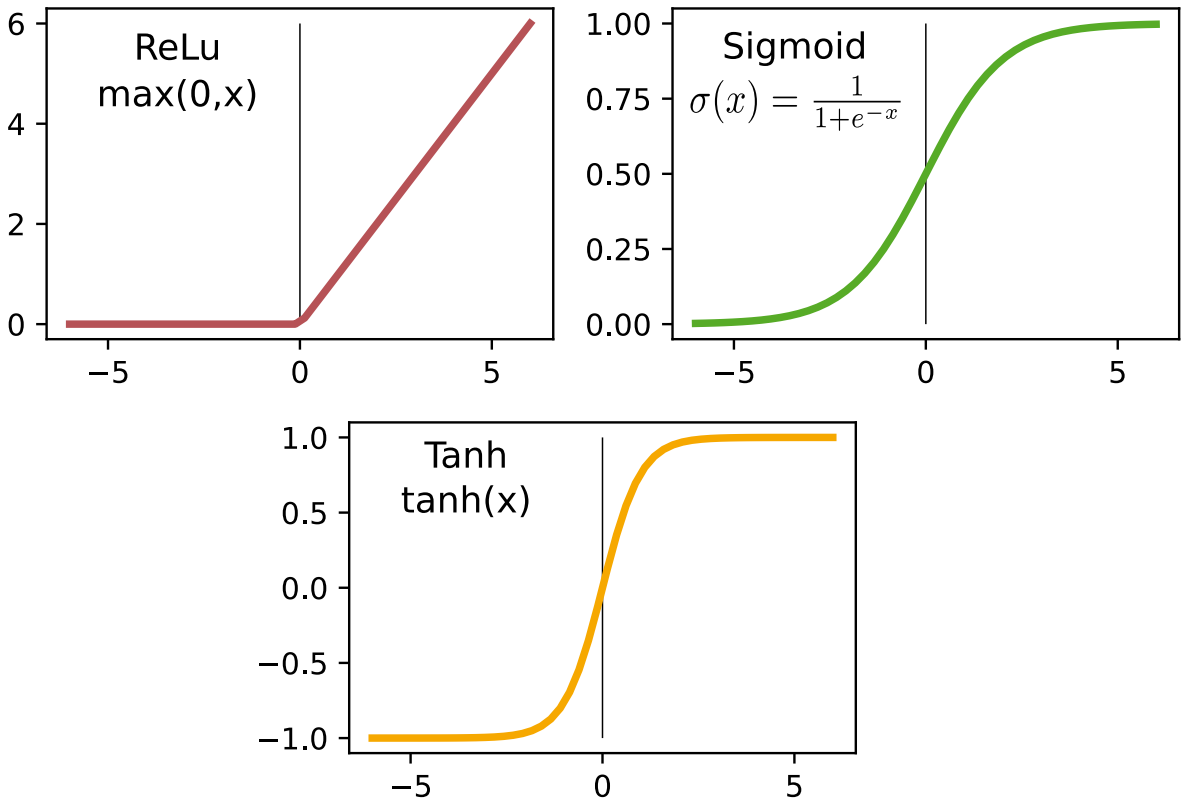


Figure 2.3: Activation functions

calculates the deviation is called *cost* or *loss* function. It is important to choose the right loss function for a model.

For multi-label classification tasks, *categorical cross-entropy* function is used. For each category, cross-entropy is calculated. The difference between the cross-entropy of training data and the model's predictions is the cost function.

$$CCE = -\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i \log(y_i) + (1-\hat{y}_i)\log(1-y_i)] \tag{2.2}$$

For regression tasks, the models are subjected to loss functions such as *mean absolute error*(MAE), *mean squared error*(MSE) and *mean squared logarithmic error*(MSLE). In MAE, the mean of absolute differences among predictions and expected results are calculated.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.3}$$

In MSE, the mean of squared differences among predictions and true outputs are calculated.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.4}$$

In MSLE, the mean of relative distances between predictions and true outputs are calculated.

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(log(y_i+1) - log(\hat{y}_i+1))^2 \tag{2.5}$$

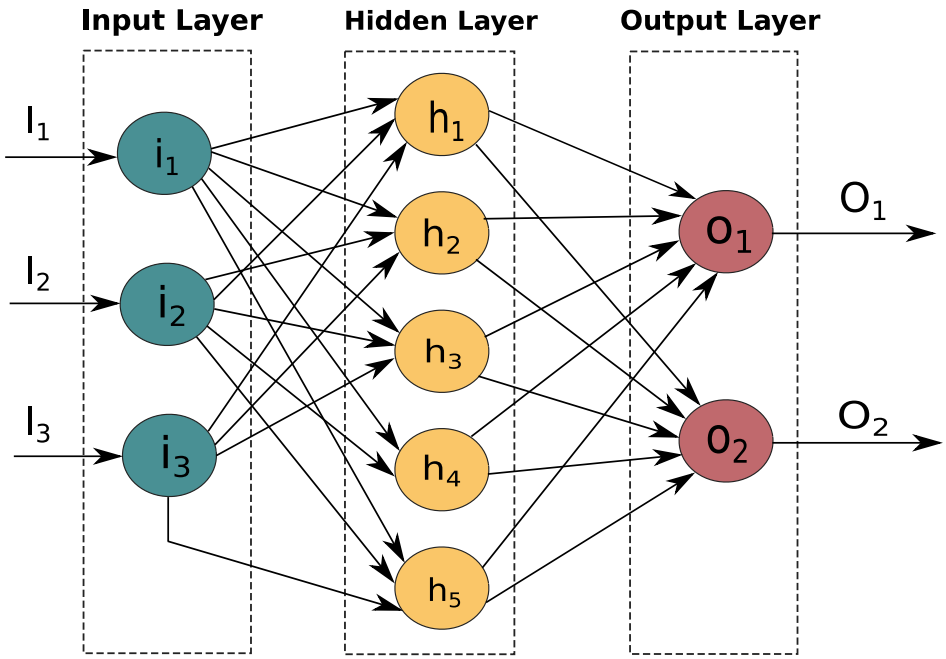


Figure 2.4: Multi layer perceptrons

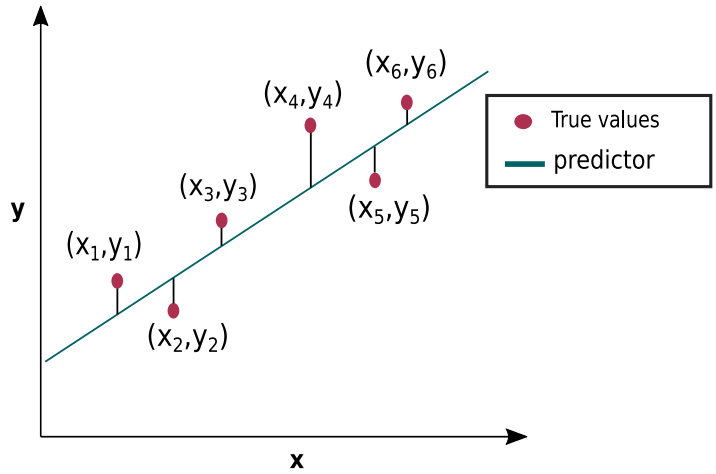


Figure 2.5: Mapping from x to y. The predictor is shown as linear line. The distance between the true values and predictor gives the loss. The sum of all the distances gives the loss function.

2.2.5 Gradient descent

Gradient descent is an optimization algorithm to minimise the cost function parameterised by a model parameter $\mathbf{w}$ in a function f. The first derivative(or gradient) gives the slope of the cost function. Hence, to minimise it, direction opposite to the gradient is chosen.

The rate at the which the gradient step reduces is given by the *learning rate*. It is one of the important parameters in training a model. It is also easily controlled by the user. Higher the learning rate, greater the step size of each gradient; possibly causing the step to miss the global minima. Lower the learning rate, more the number of steps or training cycles needed to reach the global minima. Greater care must be taken in choosing the learning rate when training a model.

2.2.6 Backpropagation

Backpropagation is the practice of fine-tuning the weights of a neural net based on the error rate (i.e. loss) obtained in the previous epoch (i.e. iteration). Proper tuning of the weights ensures lower error rates, making the model reliable by increasing its generalization. This practice is a part of model training these days.

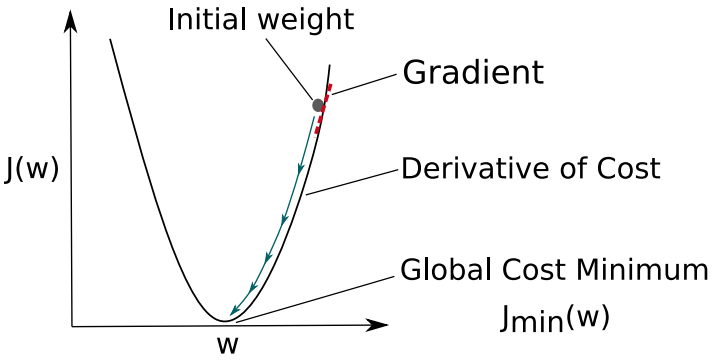


Figure 2.6: Finding the stochastic gradient descent

At the heart of backpropagation is an expression for the partial derivative $\frac{\partial C}{\partial(w)}$ of the cost function C with respect to any weight w (or bias b) in the network. The expression tells us how quickly the cost changes when we change the weights and biases. Since this method requires computation of the gradient of the error function at each iteration step, we must guarantee the continuity and differentiability of the error function. This can be achieved by using appropriate activation function such as tanh, sigmoid etc.

2.2.7 Optimizer

The loss function explains how far the predictions are compared to the true outputs in a mathematical way. During training process, certain parameters can be tweaked to help the loss function predict correct and optimised results. However, there are question such as how to change them, by how much and when?

This is exactly optmizer's function. As explained in 2.2.5, gradient descent and learning rate form the core of optimizer's functionality. *Stocastic gradient descent*(SGD) is one of the oldest techniques in which gradients for all of training examples are calculated on every pass. Hence, they are slow and require much computation power. Some of the other popular optimizers are Adam [32], Adagrad [33], RMSprop [1]. In this work, Adam is used. Adam stands for adaptive moment estimation. It is a combination of all the advantages of two other extensions of SGD – Adagrad and RMSprop. Adam is computationally efficient, straight forward to implement, invariant to diagonal rescale of the gradients, and less effort need to hyperparameters tuning.

2.2.8 Challenges in Machine learning algorithms

1. insufficient labelled data
2. poor quality data and irrelevant features
3. overfitting/underfitting a model

The first two issues can be solved if the user is careful during data collection and does preprocessing before feeding the data into the training model. However, if the training or the test data is too small, the model is subject to underfitting or overfitting. Though our aim is to reduce the error in the training

[1] RMSprop is an unpublished, adaptive learning rate method proposed by Geoff Hinton in Lecture 6e of his Coursera Class [34]

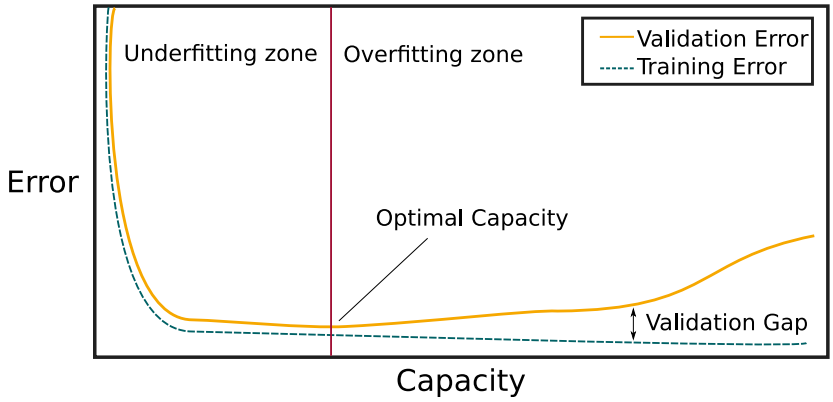


Figure 2.7: Relationship between capacity and error. Inspired from [27]

set, we also need to reduce the error in the test set. The gap between training and testing error is also an important parameter.

Underfitting occurs when the model is not able to obtain sufficiently low error value for the training set. And if the gap between training and testing error is too large, overfitting happens. The sweet spot is to stop training the model when the testing error increases while the training error decreases Left of the optimal point, the model underfits. Right of it, the model overfits. The figure 2.7 shows how the relationship between capacity and error. Validation error is the error calculated for the test set.

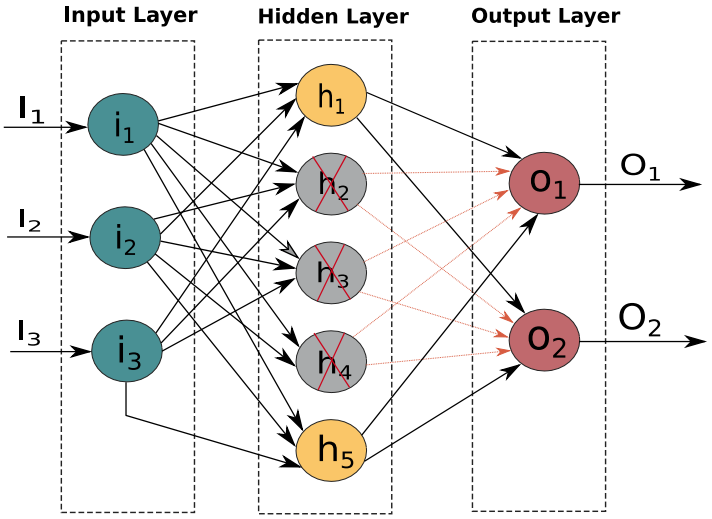


Figure 2.8: Illustrating dropout functionality

2.2.9 Regularization techniques

DNNs contain multiple non-linear hidden layers which make them easily learn complex relationships between their inputs and outputs. With a small training set, this relationship adds sampling noise that won't exist in the real-world data even if drawn from the same distribution. This leads to overfitting and several methods have been developed to reduce its effect.

1. early stopping as soon as the validation error gets worse than the training error.

2. L1 and L2 regularisation which penalises the weights [35].

3. Randomly drop units(along with their connection) from the neutral network during training [36]. Figure 2.8 illustrates how to do the random dropping of units.

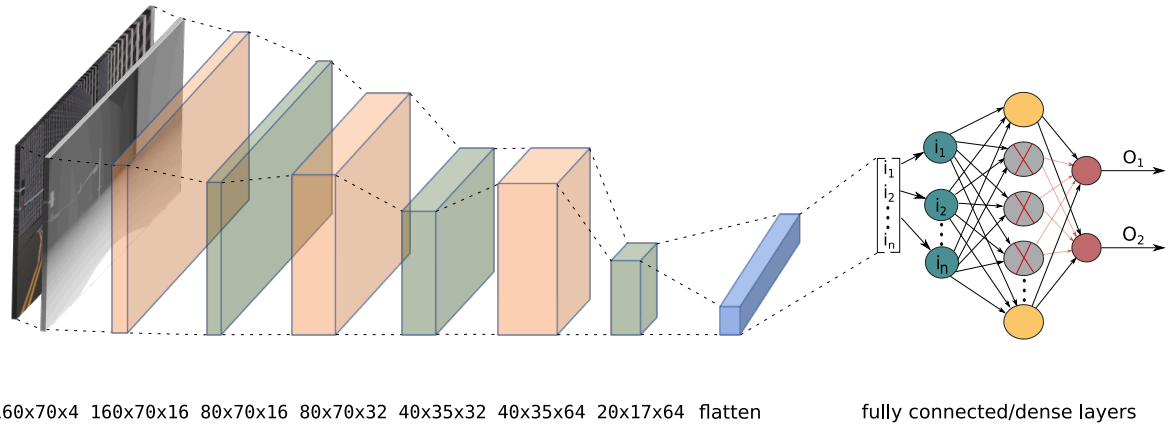


Figure 2.9: CNN architecture

2.2.10 Convolution Neutral Network - CNN

Convolutional neural network(CNN) or in short convnets is a deep learning algorithm for object recognition tasks. What makes CNN stand out for image analysis? The network takes images as inputs, reduces them into a form easier to process, without losing features which are critical for a good prediction. So not only this is important to consider while designing an architecture but also while scaling massive dataset.

Convolution Layer

The images which are taken as inputs are just n-dimensional matrix with pixel values. So a convolution operation can be easily carried on it using a filter or *kernel*. A kernel matrix is pre-defined according to the task. Usually the size of the kernel is tiny compared to that of the images which facilitates easy convolution. A *stride* is the value of a step taken by the kernel after each convolution. If stride = 1, then it is called *non-strided*. Convolution remarkably extracts the high-level features such as edges. Normally there are many Convnets in an architecture. Each layer extracts a different feature or expands on last layer's task.

If the dimensionality of the convolved feature stays the same or increased compared to the input, then it is called *same padding*. If the dimensionality is reduced, its *valid padding*. Padding is extremely useful for solving boundary conditions.

Pooling Layer

This layer is similar to convolutional layer. Its task is to decrease the computational power required to process data, usually done through reducing the dimensionality. It is, furthermore, useful to extract dominant features that are rotational and positional invariant, thus maintaining the goal of training the model.

There are two types of pooling – *max pooling* and *average pooling*. Max Pooling returns the maximum value from the portion of the image covered by the Kernel. On the other hand, average Pooling returns the average of all the values from the portion of the image covered by the Kernel. Pooling also helps in reducing the noisy pixels which sometimes skew feature extraction.

Flatten and Fully Connected Layer

The main goal for extracting features from the images is to do some task; for example - classification. So the extracted features must be converted into a form understandable for the MLP (2.2.3), which happens to be 1-dimensional vector. This is exactly the task of *flatten layer*.

MLP gets a vector as input and feeds it to a feedforward network in *fully connected layer*. Fully connected layer then outputs the necessary values depending on the task.

It is important to remember that each layer employs an activation function(2.2.2) to introduce linearity or non-linearity to the inputs.

2.2.11 Recurrent neural networks - RNN

One of the drawbacks of neural networks is that they always start from scratch; with no memory of the previous state. If a neural network has to be used for word prediction, knowledge of previous letter and word is necessary. Recurrent neural networks addresses this issue.

RNN provide the temporal dynamic behaviour. A typical RNN looks like in the figure 2.10. The left hand side shows it folded and right hand side unfolded in time. RNN, however, suffers from *long term*

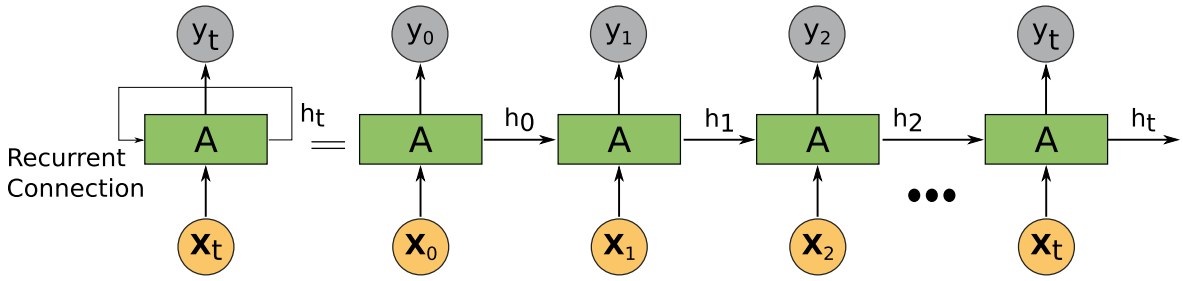


Figure 2.10: A Simple RNN

dependencies. This is well explored in [37] and [38].

2.2.12 LSTM

The shortcomings of RNN are overcome by LSTM - *Long Short Term Memory*. They are a special kind of RNN which was first introduced by Hochreiter *et al.* [39]. They remember information for long periods as their default behaviour with ease. The figure 2.11 shows how the structure of a LSTM differs from simple RNN. The LSTM employs gates and activation functions to add or delete information from the previous state.

2.3 Sensors

Deep neural networks need data – images or measurements to perform necessary tasks. These information/data are captured using sensors.

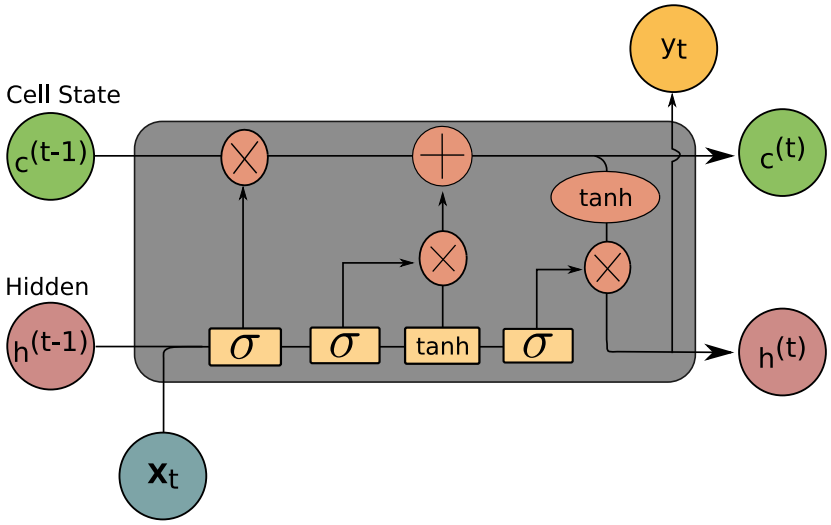


Figure 2.11: LSTM Architecture - Rolled

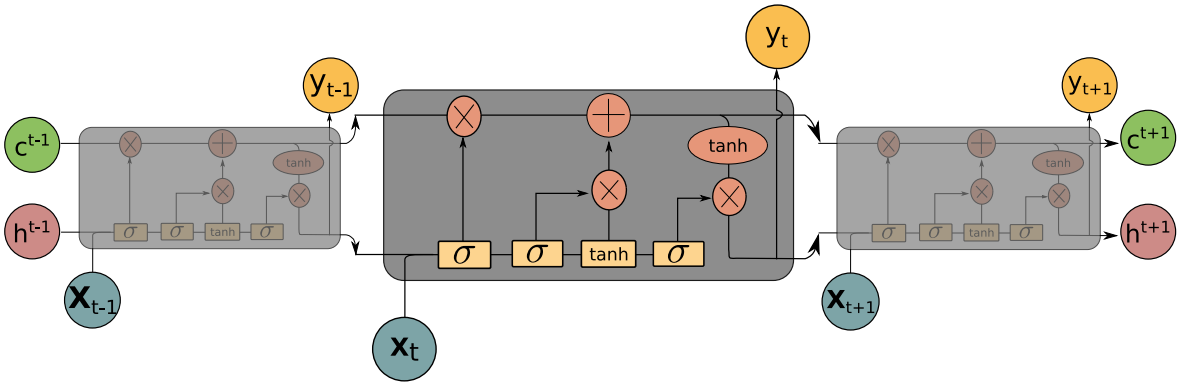


Figure 2.12: LSTM Architecture - Unrolled

2.3.1 Visual Sensors

Visual sensors are one of the commonly used sensors for image capture of the environment. Usually cameras are used.

RGB Colour Camera

A camera uses the lens to focus and captures objects. The information travels in the form of electromagnetic waves such as light. The sensor that is present behind the lens are made of photodetectors, is exposed to allow incoming light. An variable electric charge is produced depending on the intensity of the light waves. The intensity of the light waves changes with object's exposure. These charges are then quantified and stored as numerical values called *pixels*.

A pixel is generally the smallest single component of an image. Each pixel are arranged one after another in the form of matrices. So, for a resolution of "640 x 480" display, they are 640 pixels side to side and 480 from top to bottom. A colour image is captured by using a colour filter such as Bayer filter to filter out only light waves that are of RGB colour spectrum wavelength. Each pixel are then coded in *bits*. In our case we use eight bits to represent an image.

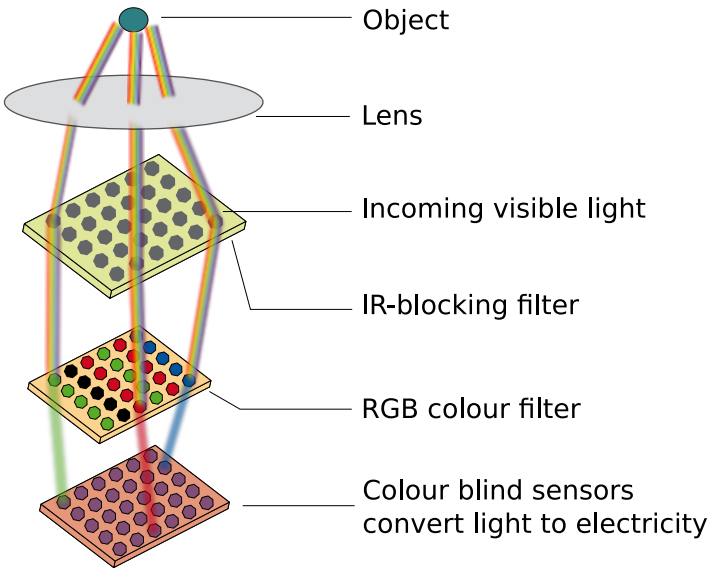


Figure 2.13: Inside RGB camera

Depth Camera

A depth camera is usually a stereo camera with two cameras. They are displaced horizontally from one another. They are used to obtain two differing views on a scene. Images are captured from these points and comparing the pixel values gives the relative position of the objects. A Depth camera

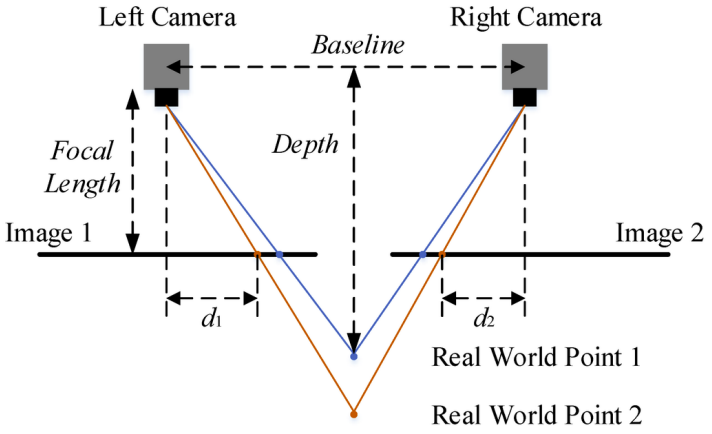


Figure 2.14: How depth sensor works. Figure redrawn using a website [40] as reference.

sensor, in our case, captures images just like a colour camera but only as grayscale images – black and white pixels. These pixels are then stored with eight bits per pixel(shades of grey). The shades on the grey-scale correspond to the depth of objects. This paper [41] shows how to combine a low resolution Time-Of-Flight (TOF) depth image camera based on Photonic Mixer Devices with two standard cameras in a stereo configuration and without accurate calibration. And this paper [42] gives the basics on stereo camera for object perception.

Segmentation Camera

RGB images are fed to CNN-DNN which groups the pixels of similar attributes using a process called image segmentation. The most common image segmentation method is thresholding. Pixels

with certain threshold are grouped together. This allows to images to have segments that may be more meaningful to analyse those segments than the whole image for relevant information.

For autonomous driving, semantic segmentation method is used. RGB images are fed to a DNN to group pixels according to user-defined tag. For example, a car is blue, pedestrian red, road boundaries white etc. This paper by Poudel *et al* [43] uses encoder-decoder architecture to do offline semantic image segmentation.

2.3.2 Measurement Sensors

These sensors are required for providing information other than visual such as telemetry data.

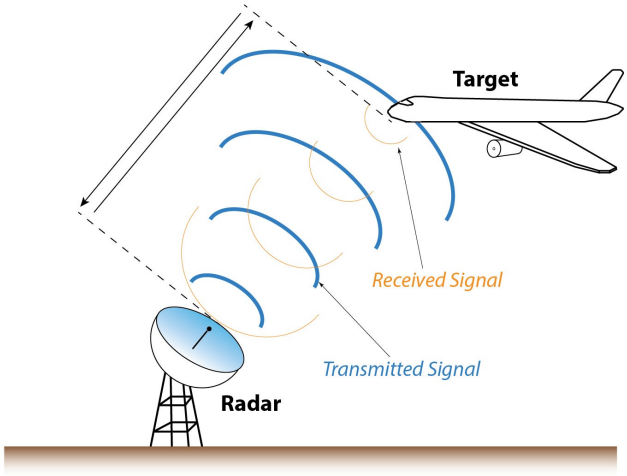


Figure 2.15: How radar works

Radar Sensor

Radar is a detection system that uses electro-magnetic waves(radio waves) to determine the range and velocity of objects. A common sensor that is used in weather forecast which is known for its long range sensing and resistance to adverse weather conditions. It consists of a transmitting antenna with a transmitter producing radio or microwave waves, a receiving antenna and receiver to process the information. Radio waves reflect off the object and return to the receiver carrying the information about object's velocity and position. If the object is moving either toward or away from the transmitter, there will be a slight change in the frequency of the radio waves due to the Doppler effect.

Three major classes of radar systems are typically employed in automotive active safety systems:

1. Short-range radar (SRR) for collision proximity warning and safety, and to support limited parking assist features.

2. Medium-range radar (MRR) (24GHz) to watch the corners of the vehicle, perform blind spot detection, observe other-vehicle lane crossover and avoid side/corner collisions.

3. Long-range radar (LRR) (77GHz) for forward-looking sensors, adaptive cruise control (ACC) and early collision detection functions.

However, in LGSVL simulator, only high-level implementation of radar is implemented. It does not use any waves, reflection or occlusion techniques. Since the simulator is aware of vehicles' position at any given moment in the map, the information is converted to depth information and stored.

Control Sensor

With this sensor, telemetry information can be collected. This is usually done by encoding the key presses in the keyboard.

2.4 Sensor/Data Fusion

To allow DNNs to make the best perception of the environment, it is necessary to fuse data from several sensors and feed that combined data into the DNN. This technique of fusing information exists for decades [44]. Often used data fusion technique is *Kalman filtering* and its variant [45] [46]. [47], [48] give a comparison in performance between using Kalman filter and LSTMs. [49] uses recurrent YOLO(LSTMs) to track objects through space and time.

For autonomous driving, RGB and depth information(RGB-D) is vital for obstacle avoidance. [25] uses data fusion to get better results for their experiment.

2.4.1 Types of Data Fusion

There are two traditional approaches to data fusion – *early fusion* and *late fusion*. In early fusion, all the sensor inputs are concatenated before being fed to the CNN. Whereas in late fusion, each sensor inputs are fed to separate convolutional layer and down the line, they are concatenated together.

These techniques can be seen in action in this [24] recently published paper from Facebook research team.

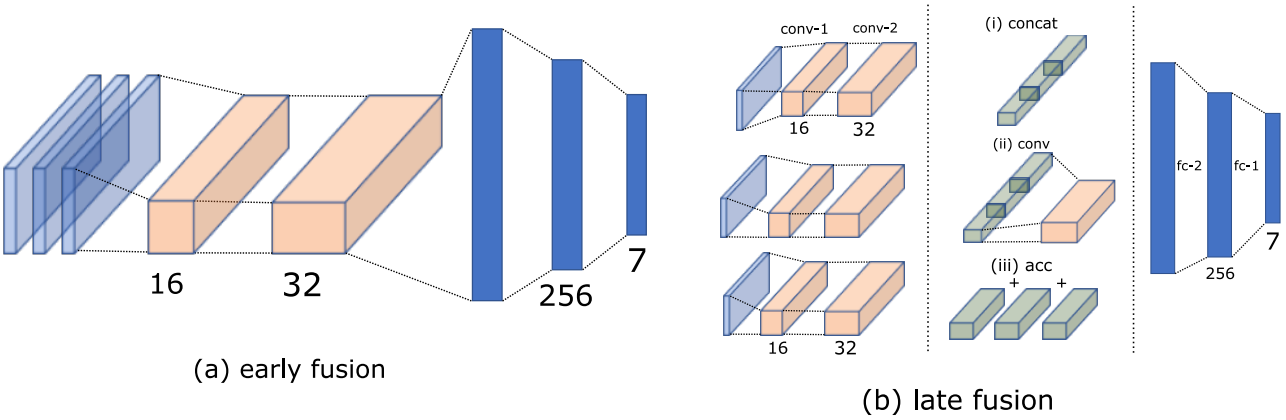


Figure 2.16: This figure is taken from this [50] paper where they describe early and late fusion architectures and also present three types of late fusion.

2.5 Machine Learning Library

To train the neural networks, we need a ML library framework to programme it. *Tensorflow* [51], *Keras* [52], *Pytorch* [53] are some of the popular ML frameworks used today. For this thesis, we

use Keras, a python high-level wrapper for tensorflow. With keras, one can easily design DNN architectures with minimal effort. All the DL techniques we discussed above are reduced to a bunch of human understandable commands. Keras has several application programming interfaces(APIs) – Models, Layers and Callbacks.

2.5.1 Models API

They are of two types – Sequential and Functional. Sequential is just stack of layers with one input and one output. Functional can handle models with non-linear topology, shared layers, and multiple inputs or outputs. Since functional API offers flexibility, it is used here. In addition to offering the overall functionality, this API has the power to implement optimizers (2.2.7), loss functions (2.2.4).

2.5.2 Layers API

The layers needed for CNN – convolution (2.2.10), pooling (2.2.10), normalization, regularization, activation (2.2.2) and time series operations with LSTMs (2.2.12) are easily implemented with this API.

2.5.3 Callbacks API

With this API, some of the overfitting challenges can be automatically avoided. Some functionalities available are early stopping (1) and ModelCheckpoint (2.2.8).

Early stopping sets an epoch parameter *n*. If the gap between training and validation loss don't improve/reduce for the next *n* defined epochs, the training is automatically stopped.

With ModelCheckpoint, the gap is monitored w.r.t a monitoring parameter; usually minimum validation loss or maximum validation accuracy. Then automatically the best model gets saved.

In order to visualise the performance of the training, *TensorBoard* class is used.

2.6 Robotic Operating System - ROS

The Robotic Operating System(ROS) [54] is a set of software libraries and tools created to help developers build robot applications. From drivers to state-of-the-art algorithms, and with powerful developer tools, ROS is a necessary set of tools for any robotics project. And its all open source.

ROS environment was first developed by Willow Garage for the PR2 robot [55]. PR2 is a humanoid robot that can navigate autonomously in a known environment. Since then, ROS is now used in all kinds of robots in various fields. With its popularity, many companies manufacture ROS compatible robots. This massively helps in integrating multiple components to communicate with each other.

ROS(ROS1), since its launch, was considered as a middleware/interface between components. There was a parent to which all the children were connected. Every child node had to go through the parent every time to discover another node. In today's expanding robotics market, this approach is outdated. This led to the development of ROS2 [56].

2.6.1 ROS2

ROS2 uses a data distribution service(DDS) for publishing and subscribing instead of custom message handler. With DDS, the transmission performance is also improved. Each node is *peer-to-peer* and

can contact other nodes efficiently. ROS2 is not simply an extension of ROS1; although some of the functionalities have been ported.

2.6.2 ROS2 concepts

In this section we will study different concepts used in the thesis.

Nodes

A node is an entity that uses ROS protocol to communicate with each other. In a ROS graph, there are networks of nodes and connections between them.

Messages

Messages are ROS data type that are used when subscribing or publishing to a topic.

Topics

A topic is named information *bus* over which nodes exchange messages. A topic usually begins with *"/"* followed by the topic name. For example, "/radar" is topic associated with radar bus. Each topic carries information of a particular message type. This message type can either be a standard or custom type.

Subscriber and Publisher

If a node subscribes to a topic, then the node is called a *subscriber*. If it publishes to a topic, then it is a *publisher*.

Both publisher and subscriber when they are initialised over a topic, a *queue size* is defined. Depending on the queue size, a topic's messages can be queued and processed as needed. The figure 2.17 shows a subscriber and publisher node exchanging data with each other.

Spins and Callbacks

In computer programming, spinning is a technique in which a process repeatedly checks to see if the condition is true. In ROS, a node is set to spin with or without a condition. This enables it to do its tasks as programmed.

Also from computer programming, callback is a function that executes at a given time. There are two types of callbacks – *blocking* and *deferred*. In ROS, deferred callback is used. It means that the callback function is invoked after a node returns something. It can be a subscriber receiving a message from its subscribed topic.

Rosbridge

We are aware that there are some non-ROS robots which would need to communicate with ROS ones. So a rosbridge [57] acts as communication API between these two. Rosbridge follows rosbridge protocol. The message transport is in JSON objects. The bridge either encodes or decodes JSON objects into appropriate ROS messages.

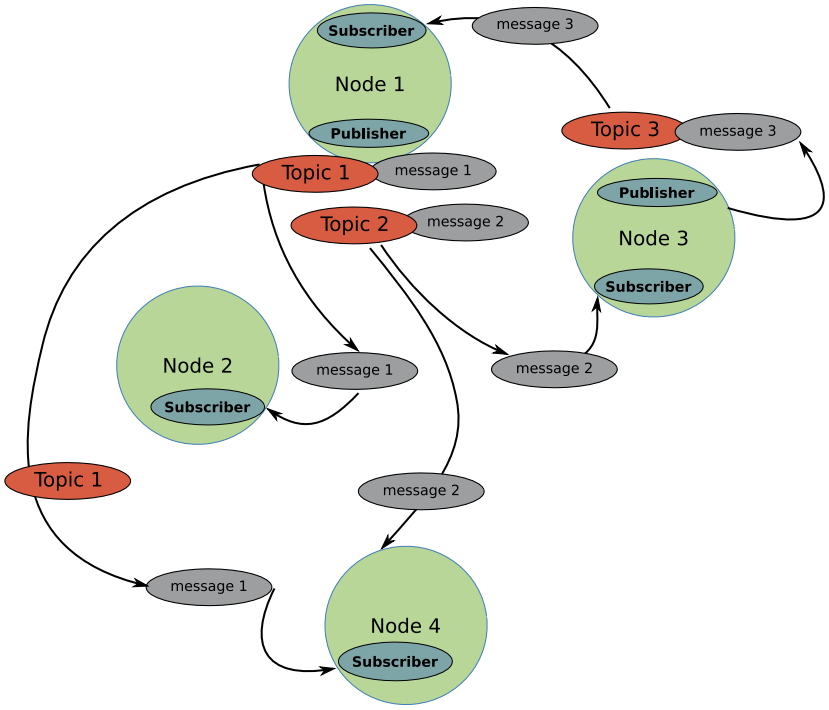


Figure 2.17: A graph showing how a publisher or subscriber node interact and exchange messages with each other through topics.

Message Filters

Since the main goal of this thesis is to do data fusion, we need to use ROS to communicate with different sensor nodes. These sensor nodes receive and transmit data at different rates. So we need a filter that can trap the received or transmitted messages, serialize them(so as to not lose data's integrity) and make it possible for storage.

With the help of a filter, all the nodes can be made to wait till every node receives the message and then invoke the callback function only once or multiple times as per design. Inside the callback, further operations can be carried out before saving.

Message_filter [58] is one such filter. It has the functionalities we are looking for, such as TimeSynchronizer, cache(a buffer to store messages while waiting for others), and slop(an extra delay parameter to TimeSynchronizer modules which defines the delay(seconds) with which the incoming messages can be synchronized.) Caution must be kept when choosing the slop value. Otherwise, the data will lose its integrity.

In the next chapter, we will see how the LGSVL [8] simulator is used.

2.7 Docker

Docker [59] is an open-source tool designed to make it easier to create, deploy, and run applications by using containers. These software containers allow developers to package an application with all the parts it needs, such as libraries and other dependencies, and deploy it as image based packages. By doing so, developers can be sure that their application will run smoothly irrespective of the client's environment. This also allows for easy debugging and development.

Docker can also be loosely considered as virtual machine(VM). But unlike a VM, rather than using a whole operating system, a docker shares the kernel of the system and ships only the applications that

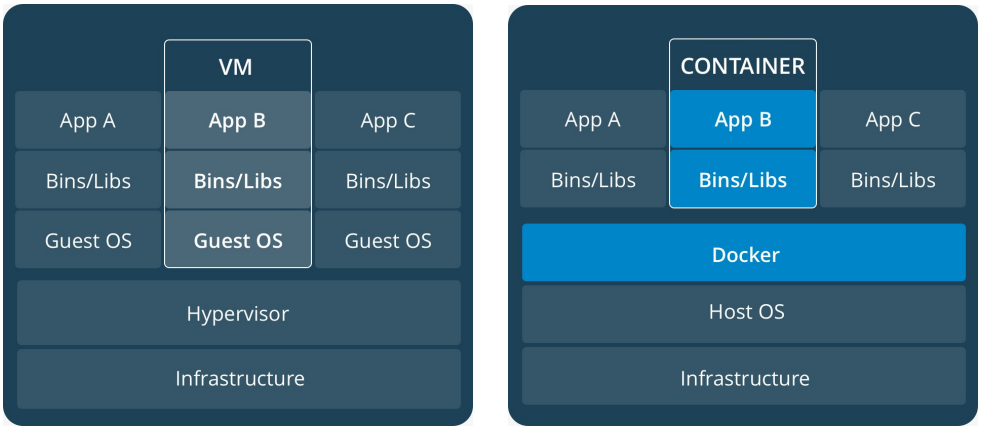


Figure 2.18: Difference between VM and Docker

are not in the host machine. Also a docker container is independent of host machine's applications. This greatly improves performance.

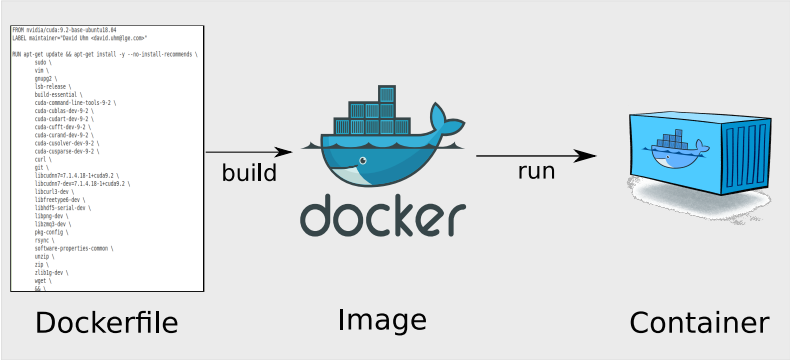


Figure 2.19: How a docker image is created

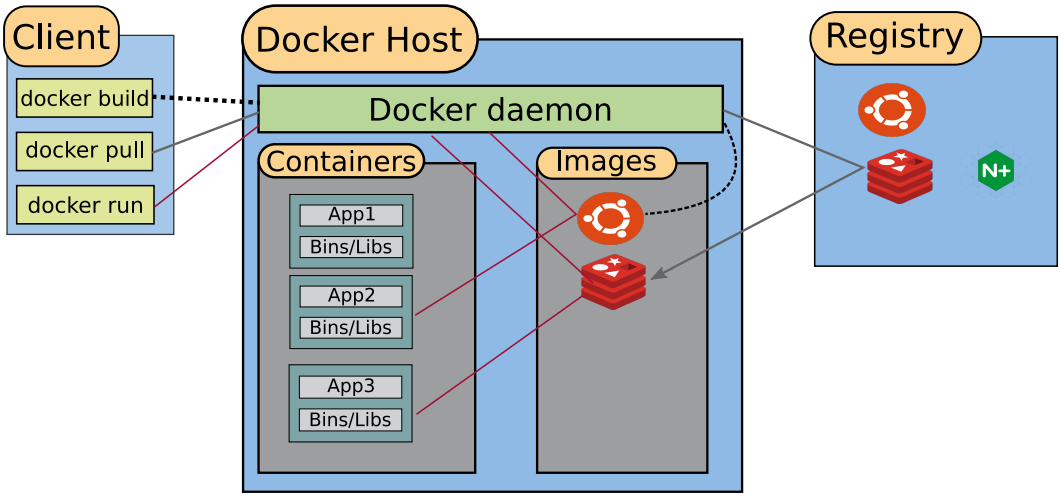


Figure 2.20: Docker Architecture

However, docker has its drawbacks. For example, building a docker image sometimes takes long to compile and consume a lot of resources. So not everyone can build an image as regularly as they wish.

In this thesis, docker containers are used for data collection from the simulator and later for evaluation of the model with the simulator.

3 Simulation and Simulator

From the beginning of autonomous driving research, simulators have played a key role in development and testing new algorithms. Simulators allow developers to quickly test their algorithms without driving real vehicles. In this chapter, we will the conditions a simulator must satisfy, and go in detail about LGSVL simulator and its development.

3.1 Need for a simulator

One of the important questions to ask before explaining about simulation is to understand why should one need a simulator to do simulation. As explained in the previous chapter (2), deep neural network(DNN) using supervised learning algorithm needs huge amount of labelled data. Since the cost of collecting that amount of data in real road vehicle is too expensive, researchers have sought the help of simulators. A simulator is an application which simulates a real-world environment, virtually. With the help of a simulator, one can collect any amount of data they wish for their project.

3.2 Conditions for a simulator

Data collection is one of the most important phases in supervised learning. So caution must be taken in choosing a simulator. A simulator must fulfil certain conditions to be qualified as a good one.

- It must have a vehicle that can move around in a virtual map.
- The vehicle must be equipped with appropriate sensors for perceiving the environment properly.
- The virtual map must try to mirror the real-world to an extent. That mean it should have proper terrain to drive around, lane marking for lane detection, other cars to mirror the real world traffic, pedestrians, and real world weather conditions.
- It must provide a medium to collect data and allow interfaces to transfer the data. It should also be able to receive data in case the user needs to validate the data collected.
- Finally and most importantly, support end-to-end, full stack simulation.

3.3 LG SVL simulator

A simulator chosen for this thesis is from LG research centre in Silicon Valley, California called LGSVL simulator. It is an open source project where the code is regularly published at Github [60]. This simulator satisfies all the conditions listed above. They provide an out-of-the-box solution which can meet the needs of developers wishing to focus on testing their autonomous vehicle algorithms. It also supports Apollo [61] and Autoware [62].

3.3.1 LGSVL simulator development

LGSVL simulator's core simulation engine is developed using the Unity game engine [63]. Unity game engine is written in C# programming language. Since a game engine inherently supports animation, the simulator is able to extend that functionality easily. In addition to Unity, also supports several libraries necessary to compute complex mathematical operations. With Unity's latest High Definition Render Pipeline(HDRP), LGSVL is able to simulate photo-realistic virtual environments that match the real world.

3.3.2 Overview of LG SVL simulator

User AD Stack

It supports user autonomous driving(AD) stack. That means a user can develop, test and verify through simulation. The user AD stack connects to LGSVL Simulator through a communication bridge interface; a bridge is selected based on the user AD stacks runtime framework. This bridge interface can use a standard protocol such ROS, ROS2 or custom one like CyberRT [61].

In addition LGSVL supports plug-in component which a user can develop and attach it to the simulator. The simulator during runtime picks up this plug-in.

Simulation Engine

As mentioned above, LGSVL uses Unity's latest HDRP game engine.

Sensor and vehicle models

It supports sensor arrangement and importantly they are customisable. The sensors are added and removed through JSON formatted text along with its parameters. These parameters include sensor type, position of the sensor, topic name, publishing rate, and in some sensors reference frame of measurement. Some of the popular sensors like camera sensors, radar and LIDAR are supported. In addition, users can add their own custom sensors as plug-in. Fig.3.1 gives a good overview of some of the sensors in action.

Vehicles provide a medium to travel the environment. Hence, vehicle dynamics is also important.

Environment and maps

An environment, in this case, virtual, is a primary component in autonomous driving simulation to provide many input to AD system. An environment affects almost all the functionalities in a AD system such as perception, prediction and tracking modules. It also affects the vehicle dynamics which is the key factor in vehicle control mechanism. Through changes in the HD map, the environment affects localization and planning modules. Finally, weather conditions such as rain, fog, night driving naturally affect the environment. So caution must be taken while design the environment.

LGSVL supports creating, editing and exporting HD maps of existing 3d virtual environment. 3D environment also defines the rules about how agents must behave such as stopping at traffic lights, giving way to priority traffic, respect lane boundaries etc.

As of writing, LGSVL supports virtual Sanfranciso city HD map. They also support smaller maps like Shalun and Cubetown.

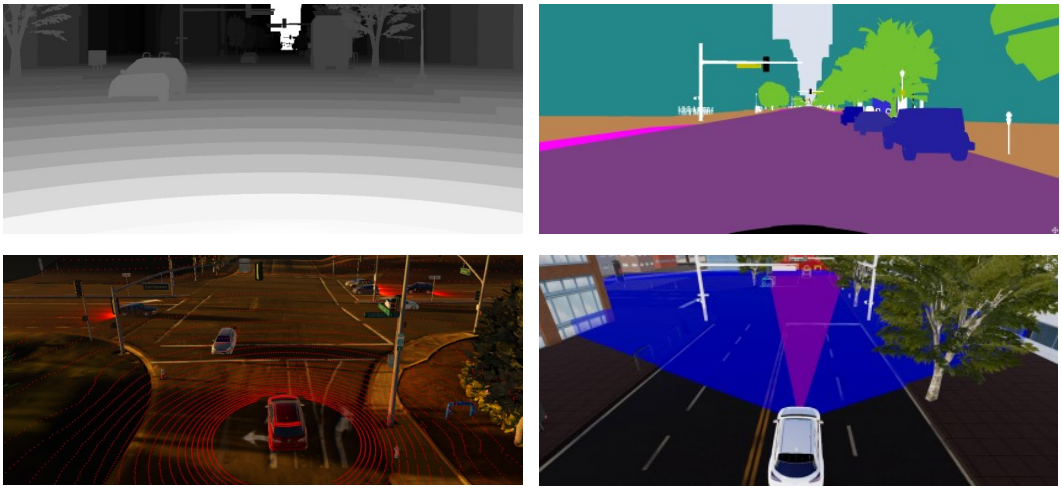


Figure 3.1: Different types of sensors in LGSVL simulator. Anticlockwise(from top): Depth camera, LiDAR, Radar(also 3D bounding boxes),and Segmentation camera

Figure 3.2: LGSVL simulator in different weather conditions

Test scenarios

Test scenarios enable users to test their AD stack by simulating in an environment and comparing and contrasting correct and expected behaviours. A lot of variables like HD maps, traffic movement behaviour and their density, time of the day, weather conditions etc. also play a role while testing. It is also possible to write scripts with the help of Python API where scenarios can be created and tested.

Thus LGSVL simulator [8] provides the best virtual environment to conduct our experiments for autonomous driving.

4 Implementation

This chapter will present the implementation of end-to-end network with its extensions. First we start with docker to set the environment. Then move on to LGSVL and ROS. From there a closed loop is achieved to collect data, preprocess, introduce neural network, implement the models, and evaluate the trained model. After achieving the basic results for the preliminary architecture, sensor fusion techniques are implemented.

4.1 Docker

Docker is an open-source platform for developing, shipping and running applications. Because docker makes installing applications hardware independent, we use docker for our tasks.

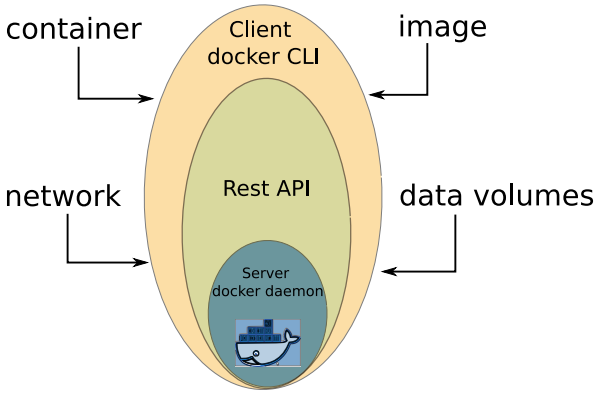


Figure 4.1: Docker Engine and its functions

A docker architecture, as shown in 2.20, consists of client, host and registry. To make all these components work, docker daemon is necessary. A daemon is a type of long-running background process. The LGSVL docker image is either pulled from the registry using *docker pull* command or built using a *dockerfile*. An image is a read-only template with instructions for creating a docker container. In our case since the image is readily available, so we pull it.

A docker container for each task can be defined. Along with a task, certain other services may need to be run along with it. *Docker compose* gives a perfect solution to manage docker applications. As seen in figure 4.2, docker-compose helps execute multiple commands. Upon execution, ROS environment are set then, a websocket with a port exposed. In our case, it is `http://localhost:9090`. The LGSVL on the other side through its web user interface(WebUI), listens on this port. So a bridge is established to flow of data.

4.2 LGSVL simulator

The LGSVL simulator [1] is developed using Unity engine which is written in C# language. The LGSVL team organises their code base[60] in such a way that it makes it easy for a beginner to learn the structure and either implement new features or change the existing ones.

[1] LGSVL simulator version till May 2020 was used for this research work. June 2020 presents new changes to how rosbridges are run which would require slight change to steps of operation.

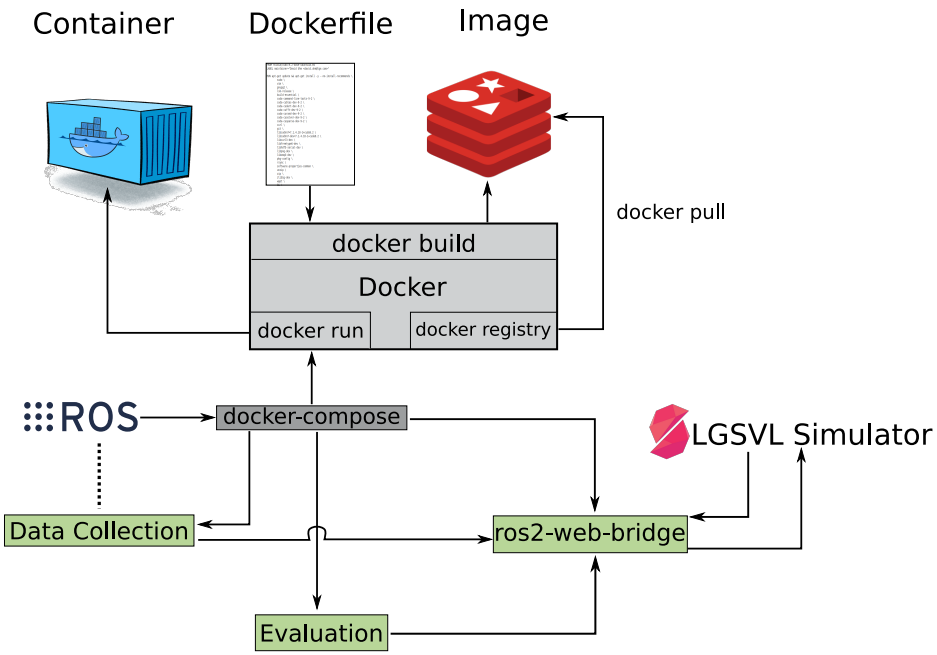


Figure 4.2: Docker and its various functions

4.2.1 Simulator configurations

The LGSVL team has developed a WebUI to help users to configure maps, vehicles and simulations. The LGSVL application connects with the WebUI through a web-socket(`http://localhost:8080`). The users are allowed to configure the simulation settings using the UI. These configurations are stored in JSON format. Some of the JSON parameters are parsed in the UI itself and some are transferred to the application using *http* protocol.

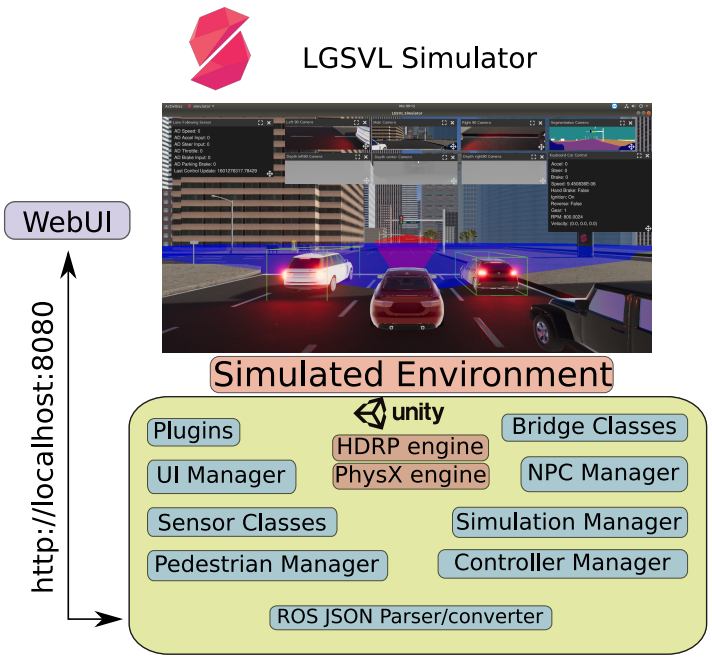


Figure 4.4: LGSVL software architecture

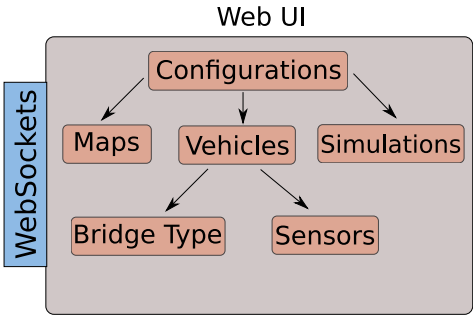


Figure 4.3: LGSVL Simulator - WebUI

Sensor parameters are defined in vehicles *tab* of the WebUI in JSON format. If a user wishes to use a colour camera sensor, then they need to use the JSON format appropriate for this sensor to the vehicle configuration. Each sensor has a *topic name*. The LGSVL application parses the JSON parameters and stores them appropriately. In order for the ROS 2 nodes to subscribe to the sensors, they would need to use the exact topic name as defined in the vehicles tab.

In our case, we use a variety of sensors:

1. RGB colour camera,
2. Depth camera,
3. Segmentation camera(uses the output of a RGB camera sensor and internally segments them according to the tags defined by the user in the sensor parameter),
4. Radar sensor.

The user is also given the flexibility to arrange/align these sensors in different constellations according to requirements using a *transform* parameter.

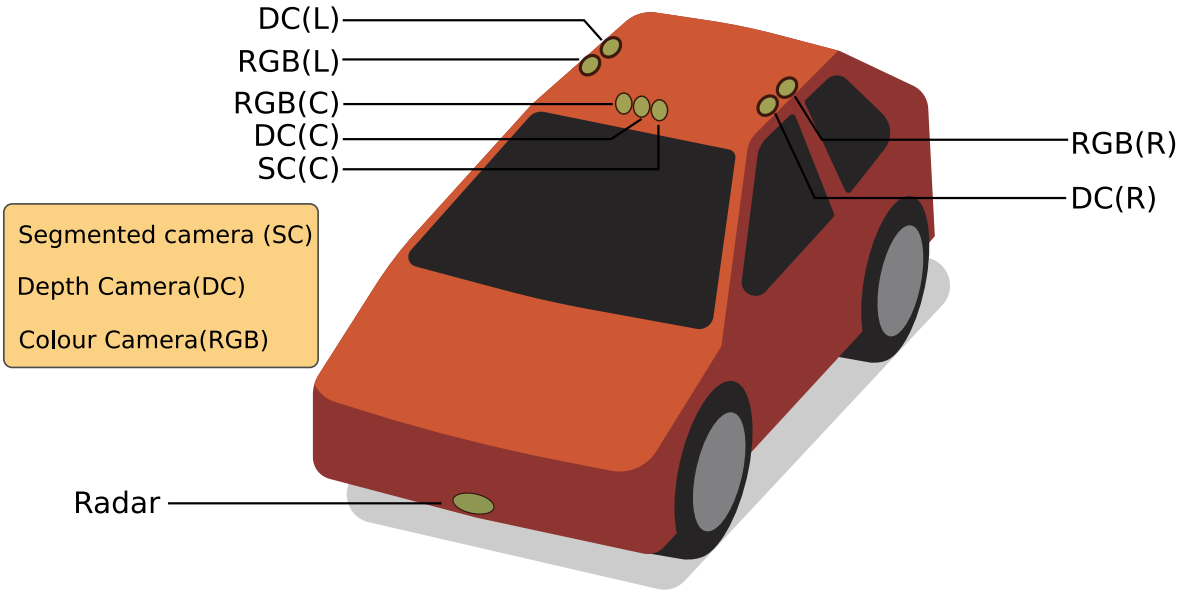


Figure 4.5: Sensor Constellation

So, we have(as seen in fig. 4.5)

1. a RGB camera placed facing ahead parallel to the ground, another on the left and right side of the car pointed an angle towards the ground,
2. a depth camera following same configuration as RGB,
3. a segmentation camera placed adjacent to the RGB front facing camera,
4. a radar sensor placed front of the car near to the hood pointing ahead.

4.2.2 Inside LGSVL simulator framework

The figure 4.4 shows some of the major functionalities inside simulator framework. When the application is first started, the WebUI is called. Natively, the app connects to UI via a websocket. Maps, vehicles, and simulation settings are communicated and initialised.

As these settings get transferred, internally, a *simulator manager* program is triggered. This is a main manager responsible for the overall functionality of the simulator. From this manager program, other managers are invoked when necessary. The *sensor classes* is responsible for coordinating with all sensors, the *NPC manager* manages traffic, signal intersections, speed limits, etc., the *controller manager* manages controls for vehicle movement. Natively, LGSVL uses Unity's *PhysX* engine for vehicle dynamics. Using Unity's *HDRP* graphics engine, the simulated environment is visualised. The *UI and Camera managers* are then responsible for the display of the environment in the application.

So, the sensor manager that oversees the sensor JSON parameters are initialised and assigned correctly, and the information from the environment using respective sensors. These sensors' data are then type cast into appropriate message data type using C# structures. Now the data resembles ROS message type. This is then passed on to *bridge classes* to be formatted as JSON and consequently publish to the subscribing ROS nodes via a bridge. The topics defined in JSON are used to publish respective sensor data.

Control Commands

The control manager is responsible for vehicle control. The steering control is defined using Unity math function as continuous float values between -1 and 1. The velocity is determined by distance travelled in the map over time. Its unit is *meter per second*. The acceleration, throttle and braking are defined as absolute values and are interconnected to one another. Acceleration ranges from -1 to 1. While accelerating, it is set as 1. Throttle is also assigned this value. While braking, the acceleration is set -1 while braking 1.

1. Steering: $a^s \in \mathbb{R} | -1 \leq a^s \leq 1$
2. Acceleration: $a^g \in \mathbb{N} | -1 \leq a^g \leq 1$
3. Throttle: $a^t \in \mathbb{N} | 0 \leq a^t \leq 1$
4. Brake: $a^b \in \mathbb{N} | 0 \leq a^b \leq 1$

4.2.3 Sensor plugin

When a user doesn't want to disturb the current setup of the simulator but rather wants to add some custom sensor to the vehicle configuration, sensor plugin can be used. A custom sensor is useful when there is a need to combine multiple standard ROS message type or define custom message types. LGSVL allows this functionality. A set of guidelines must be followed while developing the plugin. In our case, it is necessary to have a sensor plugin that would create a sensor and topics. This sensor and these topics would then be used to fetch data from the simulator, do appropriate tasks and transfer it bridge classes for transmission.

This custom plugin as shown in figure 4.6 extends the unity engine libraries just like any other sensors to read the values from the JSON definition. Upon *OnAwake*, the vehicle dynamics is initialised. Then a check whether a bridge is available is done by contacting the *bridgefactory* which holds all the bridge classes definition. In *OnBridgeSetup* method, publisher and subscriber w.r.t to the LGSVL simulator is created with topic names. At a fixed time interval, *FixedUpdate* method is invoked. Here the values obtained from vehicle dynamics are assigned appropriately. In addition if the task is to

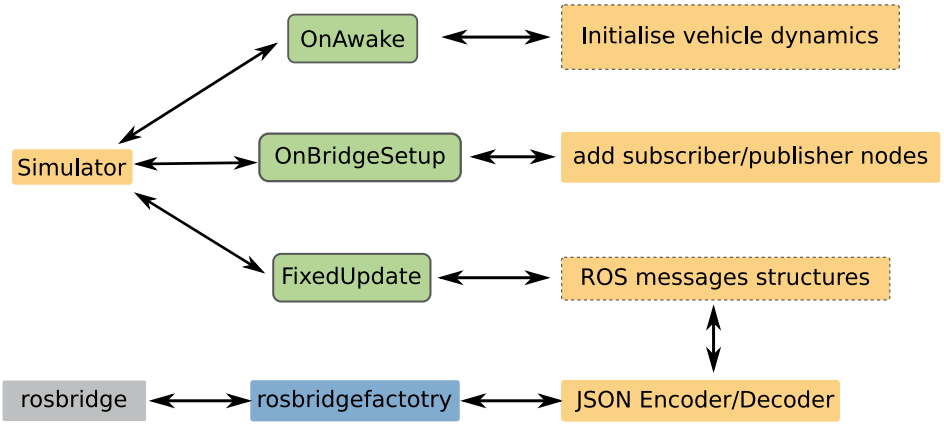


Figure 4.6: Inside sensor plugin

publish, the values such as steering, throttle, braking, velocity are converted to ROS defined JSON format through *C# ROS message structures*. If the task is subscribing, the data received through the topic such as predicted control values from the neural network model, are decoded from JSON and assigned to LGSVL variables which is reflected in the application.

LGSVL simulator is now configured to send data towards the client. In order to reach the client, as mentioned before, a rosbridge is needed. In the next section we will talk about ROS and its uses.

4.3 Data Collection Module

The figure 4.7 gives an overview of the data collection task of sensor module.

4.3.1 An overview of data gathering task

ROS, in our case, acts as an interface between simulator(server) and scripts(client). We use ROS 2 and in particular *dashing* iteration. The script's ROS 2 subscribing nodes listen to the sensors' topics(defined using JSON sensor parameters) and invoke a callback whenever they receive data. Since each sensor receives at different rates, a filter called message filters is used. With message filters, the queue size is set to a higher value for example, 1000 and a delay(in seconds) through a *slop* parameter of value 0.1 are used. This filter gathers all the subscribing nodes as one, synchronises approximately to the delay parameter and invokes just one callback. This assures that data from each listening node is present.

Inside the callback, the ROS 2 sensor data received have a header and data parts.The header part consists of the time at which the message is created and data part contains the real data. The next step would be to extract the real data. If the data are images then image processing is done. If not, the scalar values are stored in a CSV file. Then, using numpy libraries the real data is stored. The images are stored as image files. The non-image/scalar data along with corresponding image files(with filenames) are stored in a CSV file.

4.3.2 ROS web bridge

In the figure 4.8, we can see a ROS web bridge is a virtual bridge between scripts using ROS and LGSVL simulator. The *ros2-web-bridge* is just *rosbridge* which follows ROSbridge 2.0 protocol written in nodejs. It basically starts an instance that listens to an IP address and its

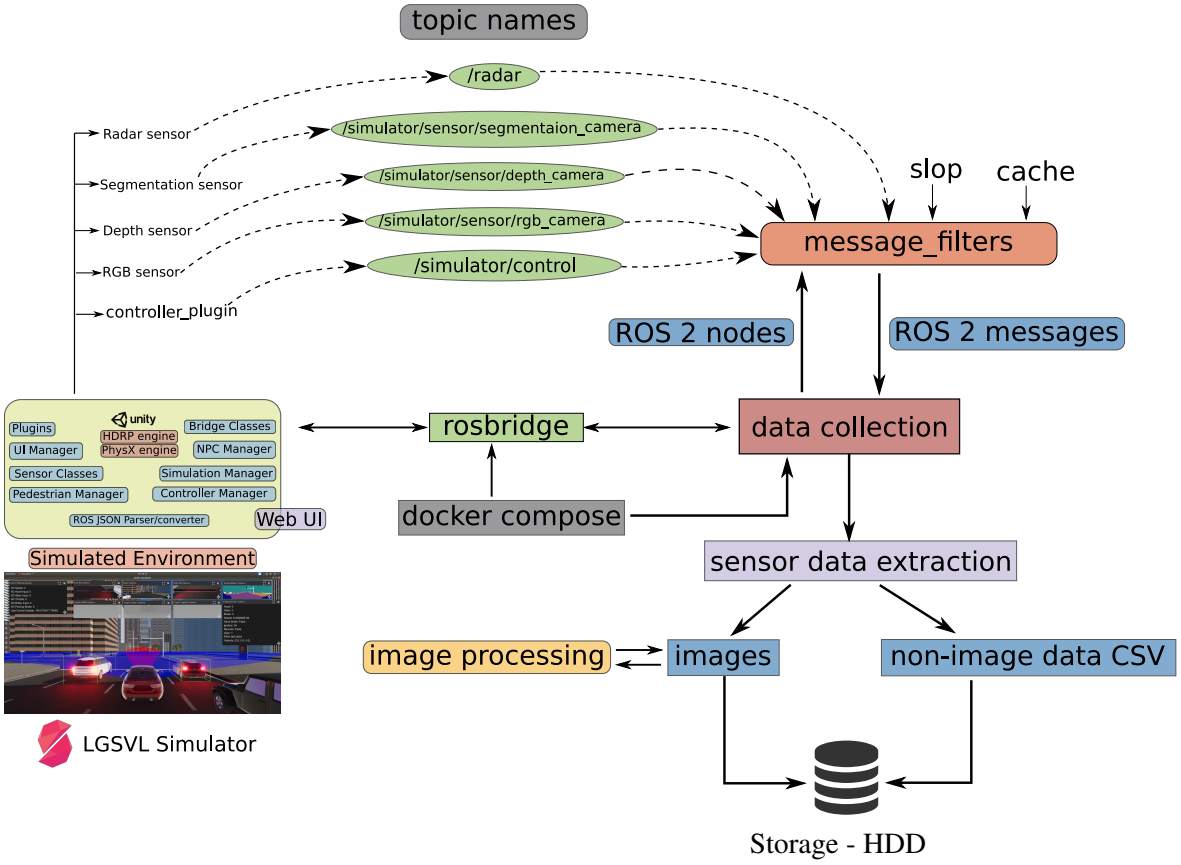


Figure 4.7: A detailed summary of data collection module

port(`http://localhost:9090`). The LGSVL on the other side(defined inside WebUI), initiates this bridge on their side at this IP address and port. Hence a bridge is created to allow flow of data.

The message transport is in JSON objects. The rosbridge server is a server which accepts websockets connections and implements the rosbridge protocol. With the help of rosbridge, data is exchanged with simulator and ROS2 nodes.

4.4 Training Module

The stored data can't be always used directly for training. Most times it must be preprocessed to user's needs and goals. Input is *X_data* and output *Y_data*. Since we use supervised learning algorithm, the output is known and labelled.

4.4.1 Preprocessing

As shown in figure 4.9, the first task in preprocessing is to select which non-image sensor *Y_data* is necessary for prediction and separate them out into a small text file. Using this file, the images which contain timestamp in the filename are fetched, manipulated using CV2 libraries, stored in arrays and saved in the form of HDF5 files [64]. The Hierarchical Data Format version 5 (HDF5), is an open source file format that supports large, complex, heterogeneous data. Within one HDF5 file, you can store a similar set of data organized in the same way that you might organize files and folders on your

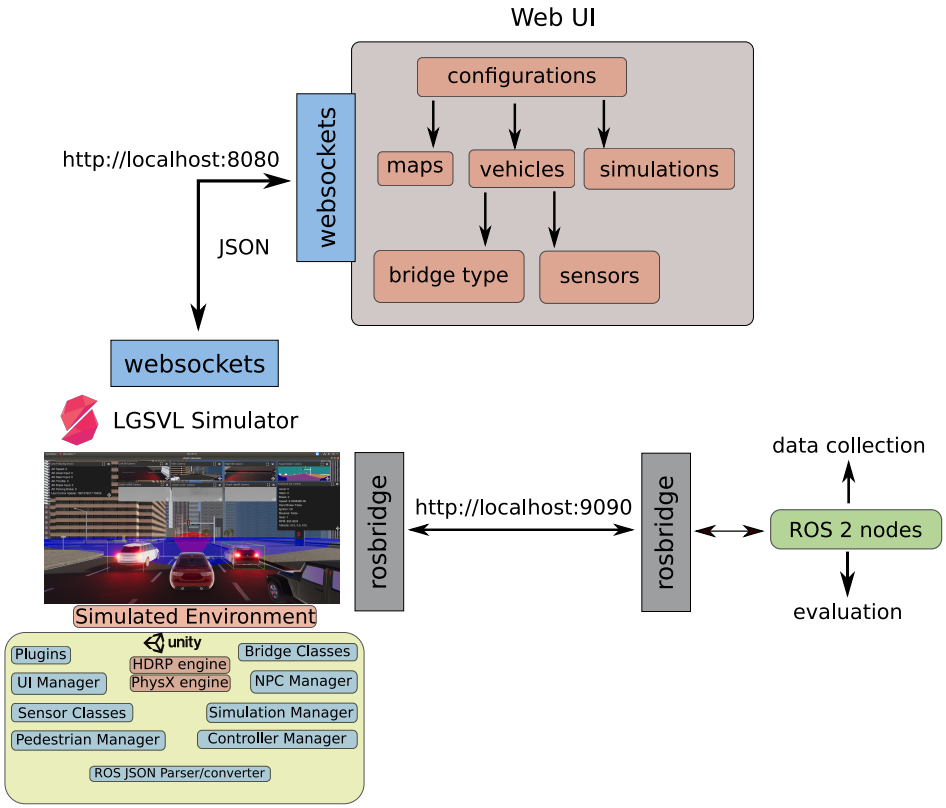


Figure 4.8: ROS2 web bridge implementation

computer. It is a compressed format and supports *data slicing* which allows only a part of the dataset to be read and not load all of them in the RAM memory.

The images in our case are read either as grayscale or RGB colour images. Then are cropped and resized to a smaller resolution such as 160x70. For grayscale image there is one channel. So the image's dimensions resemble 160x70x1 and for RGB image it has 3 channels which means the dimensions are 160x70x3.

The images from multiple viewpoints or sensors can be fused together making multi-channels. This task will be explained more in data fusion section(4.4.3).

4.4.2 LSTM

LSTM comprises of serially lined up LSTM cells which allow prediction using previous data. Since previous data require data from past, each frame image must be backtracked to a certain, defined time period. This is called *time steps*. According to the time step, the images(frames) are gathered as one and stored. So for a $time\_step = 15$, the dimensions will look like `15x70x160x1` for grayscale images and `15x70x160x3` for RGB images. In the figure 4.10, the time_step acts as a frame window. This window is moved one step to the right for every image file(frame).

So for every frame(image file), its previous 14 frames are stored along with it. However, for *Y_data* only the current frame's output is stored as we should match the current frame with its output.

Another feature we use is restricting how big is the gap between two frames. This is essential because if the consecutive frames are separated by a bigger margin, combining them for previous may lead to unknown problems. The time period we use is 1s.

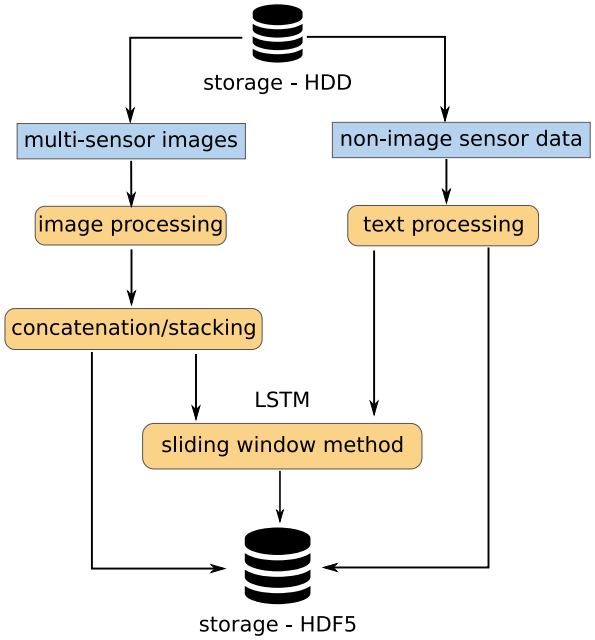


Figure 4.9: Preprocessing module

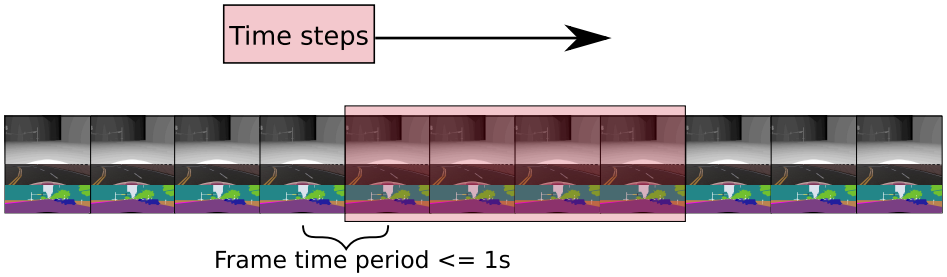


Figure 4.10: Sliding frame window implementation module

4.4.3 Datafusion

Data fusion is one of the primary goals of this thesis. As discussed in fundamentals chapter(2.4), data fusion techniques can be broadly classified into two techniques – early and late fusion. For early fusion, the images from multiple viewpoints or sensors are fused in the preprocessing stage. This fusion is accomplished either by stacking the images or concatenating them. So for example, if a grayscale and RGB images are fused/overlayed together using concatenation, then the dimensions would like `70x160x4` where 4 represents number of channels. These images are usually referred to as *multispectral images*. The figure 2.9 illustrates this approach.

Late fusion on the other hand is done during the training stage of the end-to-end work flow. Usual process involves combining(concatenating) two sources of information after one or two layers of convolution and then using the combined block to do further feature extraction and eventually prediction. Or if the source is of a different modality such as distance or velocity or number of vehicles in front of the ego car, they don't need to be feature extracted like imaged based pixel values as they are scalar values and easy of understand. Hence, it is added after the CNN is completed and the CNN outputs are made into a vector through flattening process. However, it must be remembered that late fusion increases the trainable parameters and costs on resources. The figure 4.11 illustrates one of the late fusion processes.

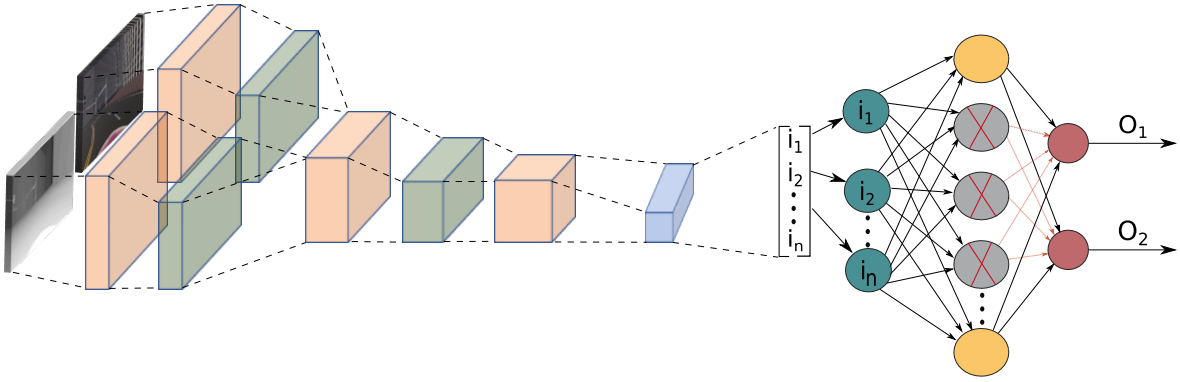


Figure 4.11: Late Fusion

4.4.4 Loading from HDF5 and splitting the dataset

The data stored in HDF5 files in preprocessing are loaded into memory as *X_data* and *Y_data* respectively. Then using scikit-learn module, the *X_data* is then split 80-20 as *X_train* and *X_test* respectively. Similarly *Y_data* as *Y_train* and *Y_test* respectively.

4.4.5 CNN and fully connected layers

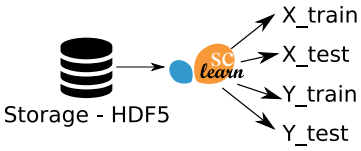


Figure 4.12: Splitting the dataset into train and test data using Sci-kit learn module.

Training a model involves designing a neural network architecture and deciding on its hyperparameters. In this thesis, CNN and dense layers are designed with appropriate activation functions, learning rate, epochs, batch size, CNN specific stride and kernel lengths, optimizer etc.

As shown in the figure 4.13, the CNN layers extract features from images through convolution. The features extracted can be stored inside feature maps. Depending on the input image dimensions, the feature maps' values are adjusted to store as many features as possible.

After extraction, these multi-dimensional data are transformed to 1-D vector through *flatten* process. Scalar, non-image sensor data can be fused at this stage. The vector 1-D data are then fed to fully connected layers which are gradually reduced till the output layer units matches the needed outputs.

The predicted output $\hat{y}$ is compared with the true *Y_train* output. The difference, *loss* is then minimised using an optimizer which does the backpropagation to adjust weights at each layer and node. The best, optimised output model is then stored in a HDF5 file.

4.5 Evaluation Module

The figure 4.14 gives an overview of the control module where the predicted control commands are evaluated. Using docker-compose the ros2-web-bridge and the evaluation are executed. Then at the LGSVL end, the same IP address and port are entered and listened in. A rosbridge is achieved.

Evaluation is basically completes the loop of end-to-end training architecture. The LGSVL simulator data are received through ROS bridge and subscriber nodes. With the help of message filters, the messages are collected. Inside the callback, the sensor data is extracted and image manipulation carried out in preprocessing phase, is repeated. The preprocessed image is then fed to the trained model.

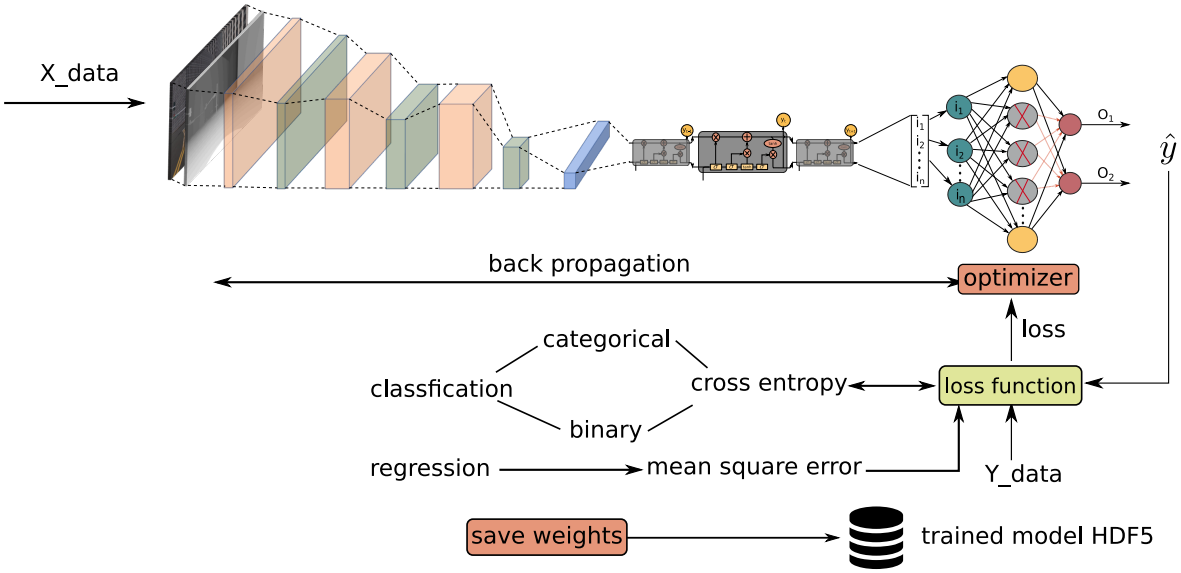


Figure 4.13: Implementation of Training module.

The model predicts the output which in our case, is control commands. These commands are then assigned and published/sent back to the simulator(sensor plugin) through rosbridge. The custom sensor plugin has a subscribing topic on the LGSVL side. The data sent through rosbridge, is picked up by appropriate data type. The predicted command behaviour is observed and evaluated using appropriate metrics. It is important to remember that, the exact steps followed in preprocessing must be repeated while evaluating. Otherwise, it will lead to inconsistent performance.

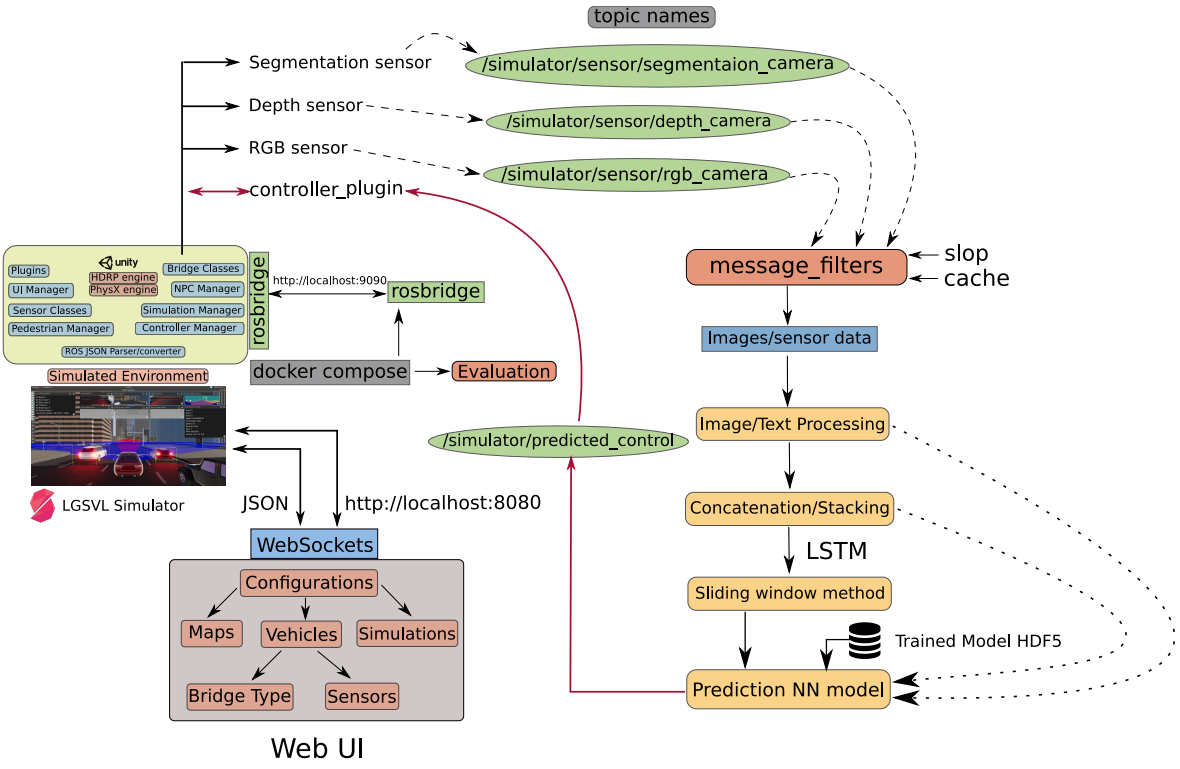


Figure 4.14: Evaluation Module

5 Evaluation

In this chapter, the workflow explained in last chapter is evaluated and results are presented.

Before showing the evaluation, it is necessary to define training and testing conditions that can be easily used by others to verify the results.

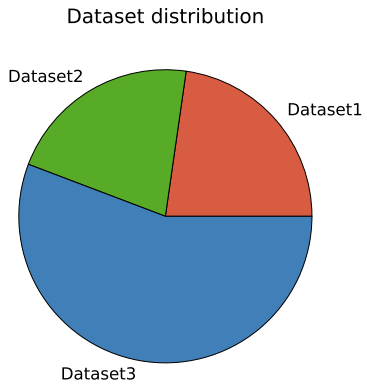


Figure 5.1: Datasets distribution

We have three datasets that can be used for training and evaluation.

1. Dataset 1 - Contains 100,000 raw data as seen in fig. 5.1. It is collected in no traffic environment, doing straight driving without any sudden turning. The data is using San Francisco map and driven during afternoon. This dataset has only data representing centre camera pointed ahead, parallel to the ground and right camera pointed to the ground at an angle 20°. The control commands include acceleration, throttle, braking, and steering angle values.

2. Dataset 2 - Also contains 100,000 raw data. It is, however, collected with traffic where the cars stop at signal intersections for a longer time than dataset 3. This dataset is also collect in San Francisco map and during afternoon. It contains a centre camera, right camera like dataset 1, left camera similar to right camera by pointing at an angle 20° to the ground, depth camera sensors placed at centre, left and right just like RGB cameras. The control commands are same as dataset 1.

3. Dataset 3 - Contains 270,000 raw data. It is collected while driving around San Francisco. About 200,000 data is collected while driving in the afternoon. About 20,000 in different weather and light conditions. About 50,000 entries are collected in a different circular circuit map called CubeTown. In addition to RGB and depth cameras distributed just as dataset 2, a segmentation camera is kept next to centre RGB camera facing forward, and a radar sensor just in front of the car near the hood also facing forward.

Hyperparameter	Value	Explanation
α	0.0001	Learning rate of the adam optimizer
Epochs	50	Number of cycles the model is trained for
β	128	Batch size
Shuffle	True	Shuffling of the data while training
CNN_FM	[24, 36, 48, 64, 64]	Values of feature maps channel of each convolutional layer
CNN_Kernel_size	(5,5) and (3,3)	Kernel or filter size for convolution
CNN_Stride	(2,2) and (1,1)	Stride parameter defining the step of the kernel
Padding	same	Keeps output image's dimensions same as input
Episode	30s	Duration of an episode in seconds

Table 5.1: Default hyperparameter setting

Evaluation setup

The table 5.1 gives an overview of all the default hyperparameter setting. Whenever something related to this setting is changed, the changes are mentioned.

While evaluating, a testing parameter *episode* is used. Each episode lasts 30 seconds. A timer is started for 30 seconds and the model is tested for collisions. If a collision happens, the time at which collision happened is noted.

As supervised learning is used, the models have to be tested/validated with unknown data to determine its capability. Hence the datasets are split 80-20. Meaning 80% is train data and 20% validation data. The optimizer *Adam* takes the 20% data to test the trained model. Training data leads to training loss and test data to validation loss.

Also till a single dataset is chosen, all datasets are of equal data entries.

5.1 Determine which datasets and best lighting conditions to test the model

All three datasets are used. The test is conducted in San Francisco map without traffic option switched ON. By varying the light conditions to morning, afternoon and evening, we observe how light influences the prediction of output(see table 5.2).

The steering angle is a continuous value ranging between -1 and 1(negative values to turn left and positive values to turn right), a continuous loss function has to be used. Because of that *mean square error*(MSE) as loss function is chosen. Only steering angle is predicted and a steady velocity of 3 meter per second is used. An episode length of 30s is used. When a collision is observed, the time of collision and the number of collisions are noted down.

time(in 24 hrs standard)	Morning	Afternoon	Evening
	7:30	15:30	18:30

Table 5.2: Time of the day

It is seen from fig. 5.2(a) that afternoon time provides the best light conditions for all the three datasets. Dataset 1 and 3 perform equally across the three lighting conditions.

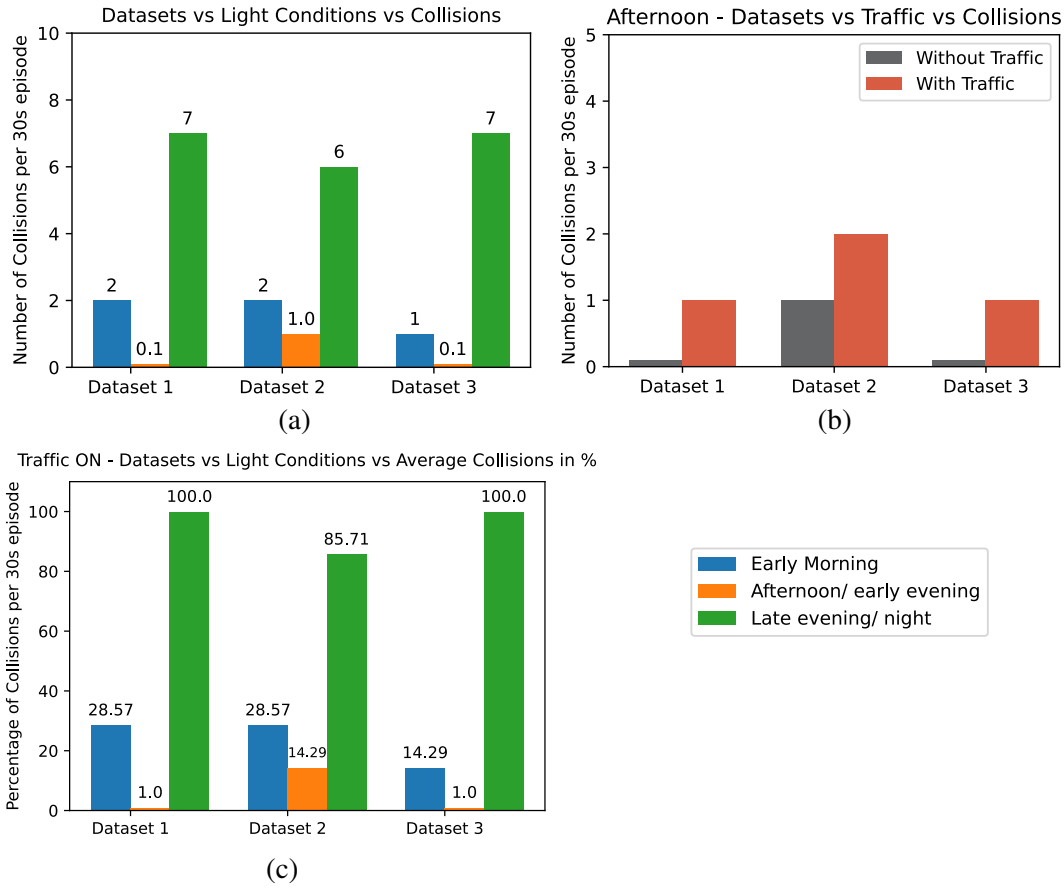


Figure 5.2: a) Datasets vs Light Conditions vs Collisions. b) Afternoon - Datasets vs Traffic vs Number of Collisions. c) Average number of collisions in percentage

If the percentage of number of collisions with traffic toggled ON, as shown in fig. 5.2(c), is calculated, dataset 3 performs the best among the datasets for morning and afternoon part of the day.

5.1.1 Datasets performance during afternoon if traffic is enabled

All three datasets are again used. The time is fixed at 15:30. The traffic is toggled ON. From fig. 5.2(b), we can observe that all three datasets do well even in traffic. However, it is surprising to see dataset 1 which had no traffic while the dataset was collected, performs remarkably well when driven in traffic.

5.1.2 Observations

1. All 3 datasets do well at afternoon time of the day.

2. Predicting only steering angle with MSE as loss function works as seen from 5.2.

3. Dataset 2 shows higher number of collisions. So it is better to avoid for further analysis.

5.2 Acceleration - Determine which activation and loss functions to use

5.2.1 Tanh as activation and MSE as loss functions

Since acceleration and steering values in LGSVL range from *-1* to *1*, *tanh* activation function is selected as the output dense layer activation function.

A set of criteria are listed and the trained model is evaluated based on these conditions. From table 5.3 it can be observed that dataset 1 outperforms dataset 3 in most of the conditions. Dataset 1 retains good steering control at high speeds and turning, but acceleration skews steering angle when traffic is switched ON. Dataset 3 when evaluated stops completely after moving a few metres. This causes difficulty in evaluating according to the criteria. Hence it is assumed that this dataset fails to meet the criteria.

One of the reasons as to why dataset 3 fails could be because the losses are much higher than dataset 1 and the validation loss starts to overfit too quickly in the training(see fig 5.3).

Criteria(Tanh/MSE). Rating 1 to 5. 1 being the lowest.	Dataset 1	Dataset 3
Lane keeping/Drive straight	5	1
Gradual acceleration increase	4	1
Smooth braking behaviour observed	4	1
Smooth steering control at high speed(10m/s)	4	1
Smooth steering control at turnings	3	1
Detects traffic as dynamic objects	5	1
Navigates traffic smoothly	2	1
Stops at random places	5	5

Table 5.3: Tanh/MSE - How the model evaluates to different criteria

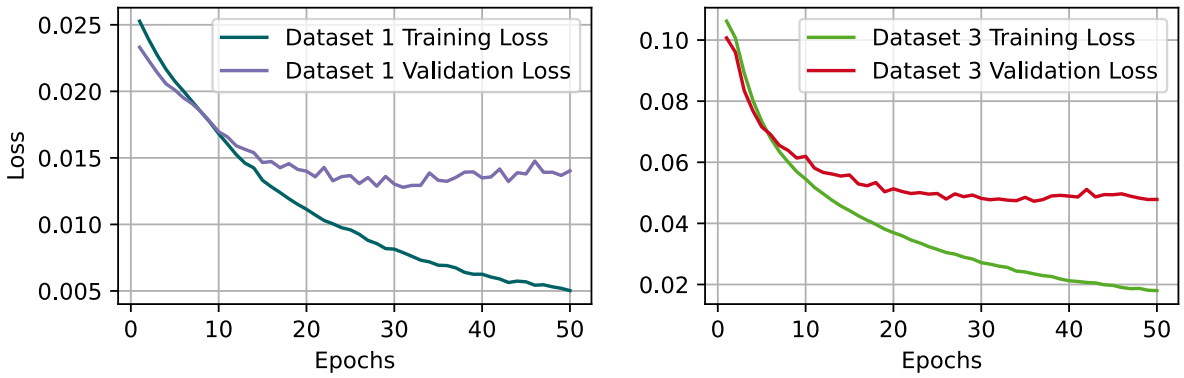


Figure 5.3: Datasets 1 vs 3 - Acceleration and Steering using Tanh activation and MSE loss functions.

5.2.2 Sigmoid as activation and MSE as loss functions

The acceleration values are split into positive and negative values. Instead of negative values an another variable we will call as *braking* is introduced. Negative acceleration values mean braking is active. Using this knowledge, sigmoid as activation function and mean square error as loss function, a training is conducted for both datasets 1 and 3.

Criteria similar to table 5.3 are put to test with this activation and loss function. As seen in table 5.4, both datasets fail to meet the conditions. During evaluation, both datasets trained models, accelerate to a huge velocity such as 40m/s and the steering cannot keep up, resulting in collisions.

Looking into their losses (see fig. 5.4), don't really explain much. Though this needs more investigation, for now we assume this activation or loss function is not suitable. 5.3.

Criteria(Sigmoid/MSE). Rating 1 to 5. 1 being the lowest	Dataset 1	Dataset 3
Lane keeping/Drive straight	1	1
Gradual acceleration increase	1	1
Smooth braking behaviour observed	1	1
Smooth steering control at high speed(10m/s)	1	1
Smooth steering control at turnings	1	1
Detects traffic as dynamic objects	1	1
Navigates traffic smoothly	1	1
Stops at random places	5	5

Table 5.4: Sigmoid/MSE - How the model evaluates to different criteria

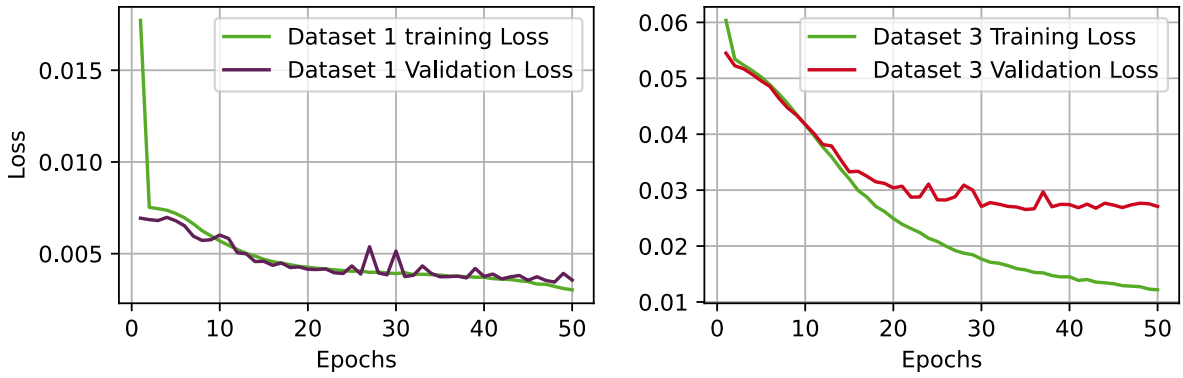


Figure 5.4: Datasets 1 vs 3 - Acceleration and Steering using Sigmoid activation and MSE loss functions.

5.2.3 Softmax as activation and Binary crossentropy as loss functions

Both tanh and sigmoid activation functions couldn't give stable results to continue pursuing with those parameters. Hence it is necessary to consider other functions which may suit our needs.

Since acceleration are basically two discrete values, it would be worthy to try the training as classification task. The goal of a classification task model is to classify to which category the prediction belongs to. In our case, acceleration or braking. Hence *softmax* activation function is needed. As loss function *binary crossentropy* is used to classify as binary classes. Of course for steering angle, being continuous, MSE is preferred.

From table 5.5, both datasets don't do well. Dataset 1 doesn't start at all as braking class dominates the evaluation whereas dataset 3 starts to move the car but resorts to brake indefinitely after a few metres of driving. This behaviour necessitates further analysis into the datasets.

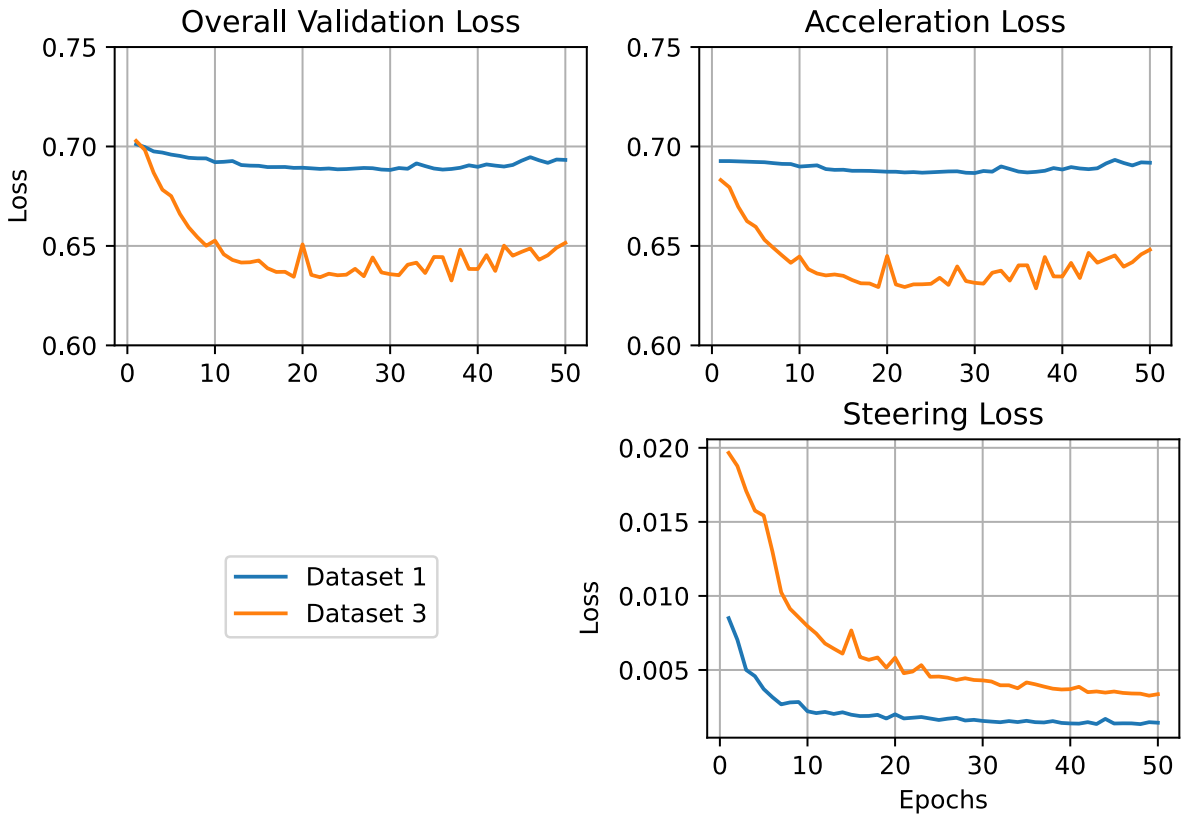


Figure 5.5: Dataset 1 vs 3 Validation loss - Binary crossentropy and Softmax functions

Control commands distribution

When the datasets are analysed for patterns of different states – acceleration and braking, distribution chart(fig 5.6) reveals that in addition to acceleration and braking states, there is a third state called *no action* where the vehicle does absolutely nothing. In fact this state dominates in both datasets. Because of this, the binary crossentropy obviously fails to meet the conditions.

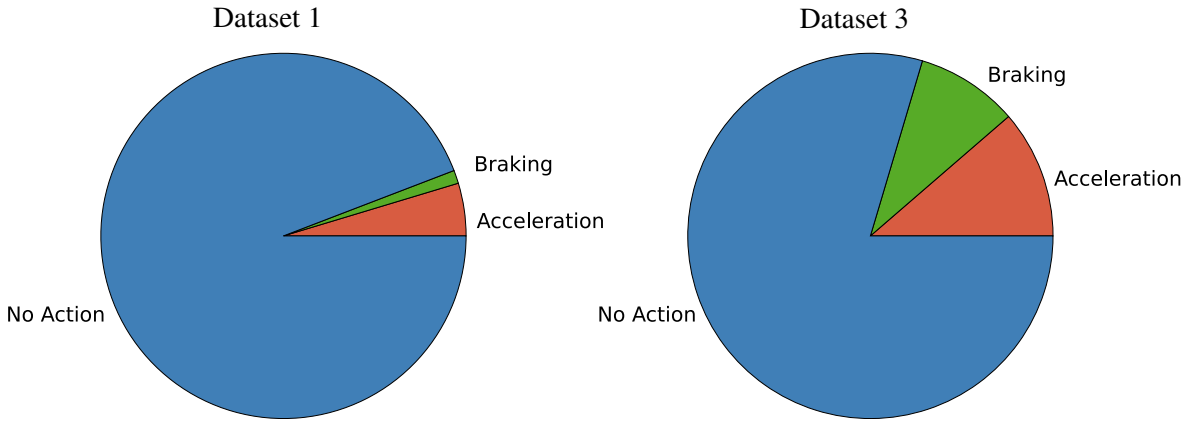


Figure 5.6: Datasets 1 vs 3 control commands distribution

Criteria(Softmax/Binary crossentropy). Rating 1 to 5. 1 being the lowest	Dataset 1	Dataset 3
Lane keeping/Drive straight	1	4
Gradual acceleration increase	1	1
Smooth braking behaviour observed	1	1
Smooth steering control at high speed(10m/s)	1	1
Smooth steering control at turnings	1	1
Detects traffic as dynamic objects	1	1
Navigates traffic smoothly	1	1
Stops at random places	1	1

Table 5.5: Softmax/Binary crossentropy - How the model evaluates to different criteria

5.2.4 Softmax as activation and Categorical crossentropy loss functions

So now we know that this dominant state *no action* needs a separate label if classification task has to be continued. After creating the label, we would then have three labels – acceleration, braking and no action. Hence, we use a new classification loss function called *categorical crossentropy*. This loss function classifies model into each category.

Criteria(Softmax/Categorical crossentropy). Rating 1 to 5. 1 being the lowest	Dataset 1	Dataset 3
Lane keeping/Drive straight	1	3
Gradual acceleration increase	3	3
Smooth braking behaviour observed	1	3
Smooth steering control at high speed(10m/s)	1	1
Smooth steering control at turnings	1	1
Avoids colliding into static objects	1	1
Detects traffic as dynamic objects	5	5
Navigates traffic smoothly	1	1
Stops at random places	5	5

Table 5.6: Softmax/Categorical crossentropy - How the model evaluates to different criteria

From table 5.6, dataset 1 fails in most conditions and dataset 3 though performs well in 4 out of 9 conditions, shows bad steering behaviour at traffic or high speeds. If the acceleration is controlled manually, the model reacts better and adapts itself.

Since no action state dominates, it is necessary to continue as a classification task. Also dataset 1 has only a small portion for acceleration and even smaller for braking. Accordingly, dataset 3 is collected with an attempt to increase acceleration and braking values share. However, since the vehicle most times has to drive straight, *no action* state even dominates in dataset 3.

The important condition in a classification task is to have balanced classes. Unfortunately in our case, this balance is not achieved. With this limitation, the evaluation is carried forward.

5.2.5 Observations

1. Steering angle uses tanh activation and MSE loss functions
2. Classify acceleration prediction as classification task which means softmax as activation.

3. Since acceleration carries three states, categorical crossentropy as loss function.

4. Dataset 3 has more of acceleration and braking states than dataset 1. So dataset 3 is preferred.

5.3 Predicting acceleration - categorical crossentropy

Now that the basic criteria for training is fixed such as which dataset, activation, and loss functions, we can move ahead and optimise the predicted classes by tuning the neutral network.

LSTM vs Non-LSTM

Before going into tuning the neural network, it is necessary to tell that acceleration prediction needs temporal information; meaning decision to drive slower or faster depends on the previous, historical frames. LSTM is used for this purpose. When non-LSTM model is used to predict acceleration, it only predicts for the current frame(doesn't provide past frames information) which often results in vehicle being stationary. For our setup, we choose a $timestep = 15$. That means acceleration of current time

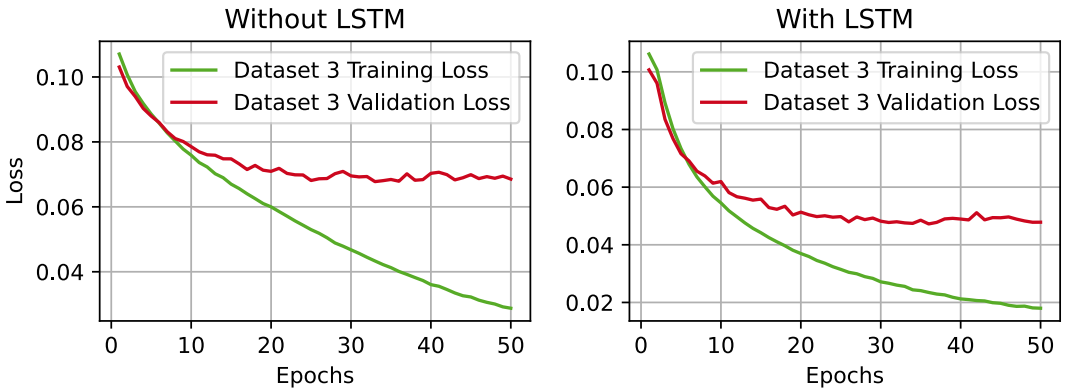


Figure 5.7: Datasets 3 - No LSTM vs LSTM comparison

frame is predicted using previous 14 time frames.

Determining the optimal LSTM output units

In Keras, the LSTM units refer to the dimension of hidden state vector h that is the state output from RNN cell. The table 5.7, compares different unit values and the number of trainable parameters(weight that can be trained during backpropagation) at this LSTM layer. In our case, it means that for a time series of 15, there will be 15 *cell states*, 15 *hidden states*, and 15 *outputs* each of vector size defined by the units in the table such as 20, 60 or 100.

Upon evaluation with these different units, 100 output units though has the highest trainable parameters for this layer alone, retains more information needed for training the model. Hence, a LSTM unit of 100 is chosen.

5.3.1 Basic Model

For training, a model as shown in fig. 5.8, is designed and its result is seen in fig. 5.9. Interestingly, the training loss curve *follows* the classification loss as it dominates the model. Steering loss however, after epoch 32 starts to increase. Sure enough, upon evaluation, the steering is all over the place and acceleration is not stable at all, resulting in many collisions as shown in table 5.6 (dataset 3 column).

LSTM Output Units	Trainable Parameters(ca.)	Processing time needed
20	20000	1hr 44m
60	61000	1hr 42m
100	434000	1hr 40m

Table 5.7: LSTM Output Units vs Trainable Parameters vs Training time

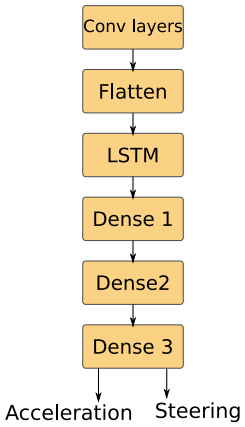


Figure 5.8: Basic model

The cause for this behaviour is investigated and it is found that different losses have different magnitudes. By forcing the neural network to learn/train both classification and regression losses from a same dense layer causes instability in learning.

5.3.2 Splitting at the dense layers

To alleviate some of the burden the second dense layer(dense 2) is split into two separate dense layers; one for classification outputs and other for steering as show in fig. 5.10. The result 5.11, stops the strange steering loss increase. Upon evaluation, this model shows better steering control but still the acceleration is not stable or consistent as shown in table 5.8.

Criteria(Softmax/Categorical crossentropy). Rating 1 to 5. 1 being the lowest	Dataset 3
Lane keeping/Drive straight	4
Gradual acceleration increase	3
Smooth braking behaviour observed	3
Smooth steering control at high speed(10m/s)	2
Smooth steering control at turnings	1
Avoids colliding into static objects	1
Detects vehicles as dynamic objects	5
Navigates traffic smoothly	3
Stops at random places	5

Table 5.8: Separate dense layers - How the model evaluates to different criteria

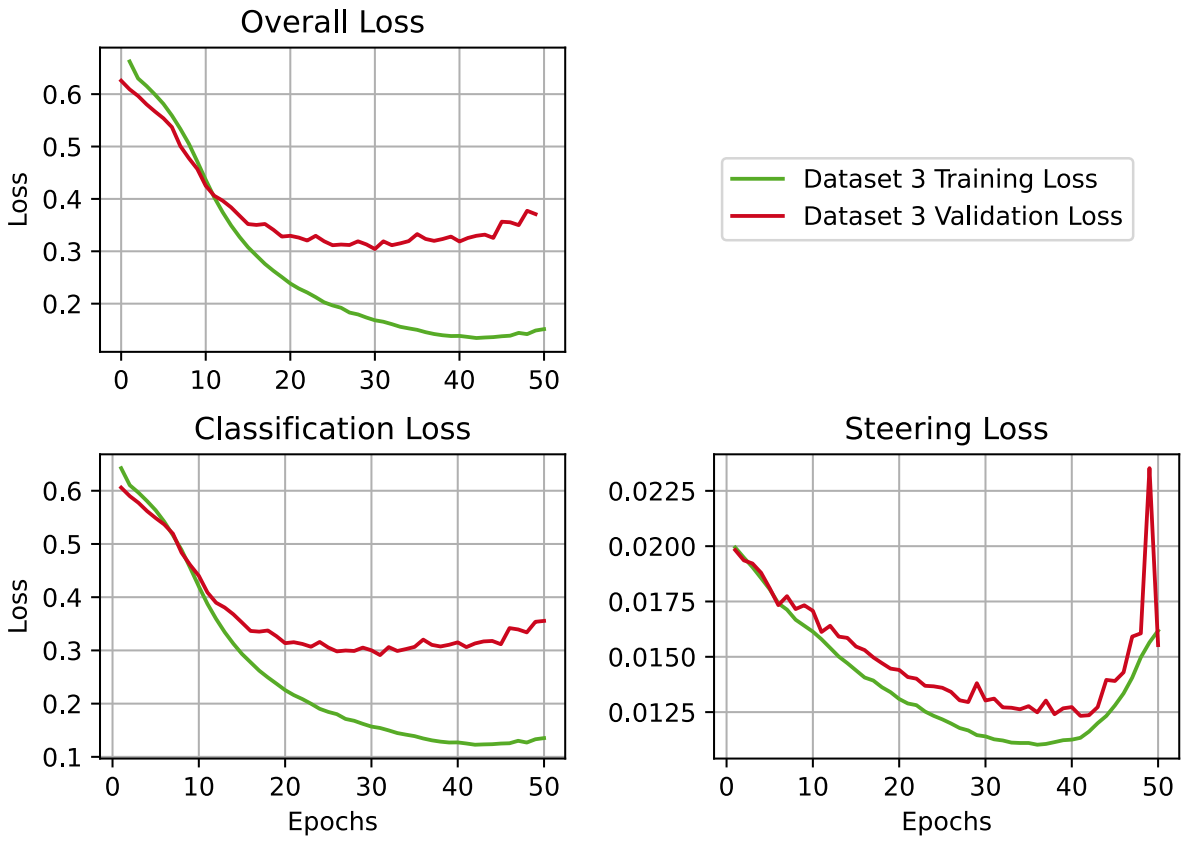


Figure 5.9: Basic model

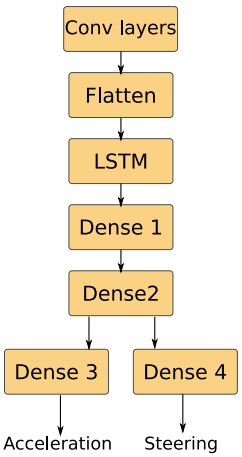


Figure 5.10: Split at the second dense layer

5.3.3 Splitting at the LSTM layers

Continuing the theme of tuning the network, the model is split further at LSTM layer as shown in fig. 5.12. The main aim here is to see if steering control improves and is stable for considerable acceleration prediction. From fig. 5.13, the steering loss gets a marginal gain. Still at evaluation, the trained models stops at random places, and steering control is not stable at higher acceleration predicted values. Hence the predicted acceleration value is reduced by 50-70% and fed to the controller. Sure enough the vehicle exhibits stable movements as seen from table 5.9.

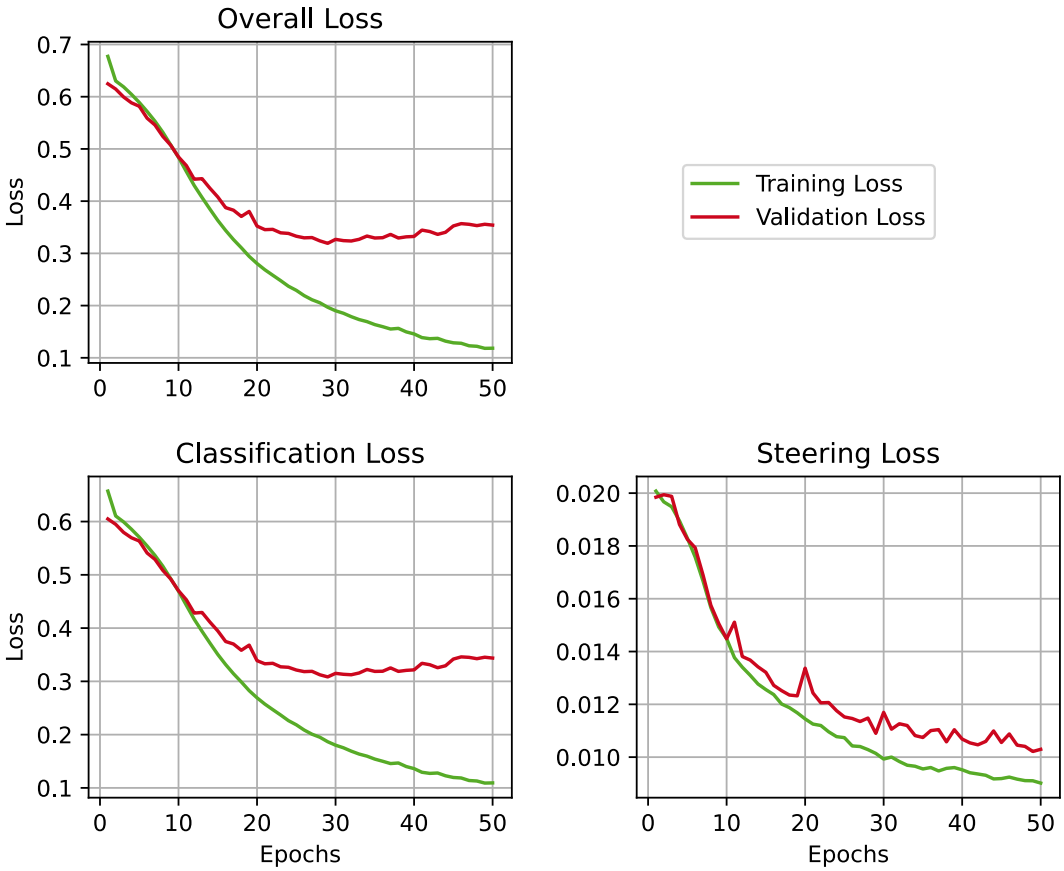


Figure 5.11: Separate dense layers for classification and steering

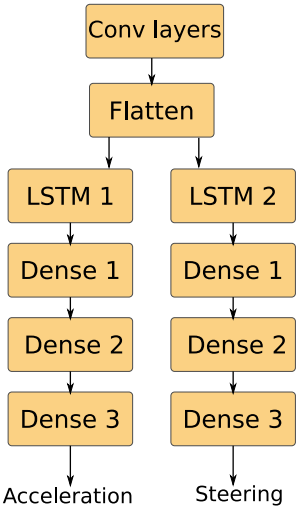


Figure 5.12: Split at the LSTM layer

5.3.4 Using two different NN for acceleration and Steering

It can be deduced that optimised weights play an important role on how it influences steering and acceleration prediction. The model is now split as two different neural networks(NN) as seen in fig. 5.14. Though the result 5.15 looks similar to 5.13, the model predicts stable, consistent acceleration values.

From table 5.10, it even exhibits occasional turning behaviours at junctions. When it is exposed to

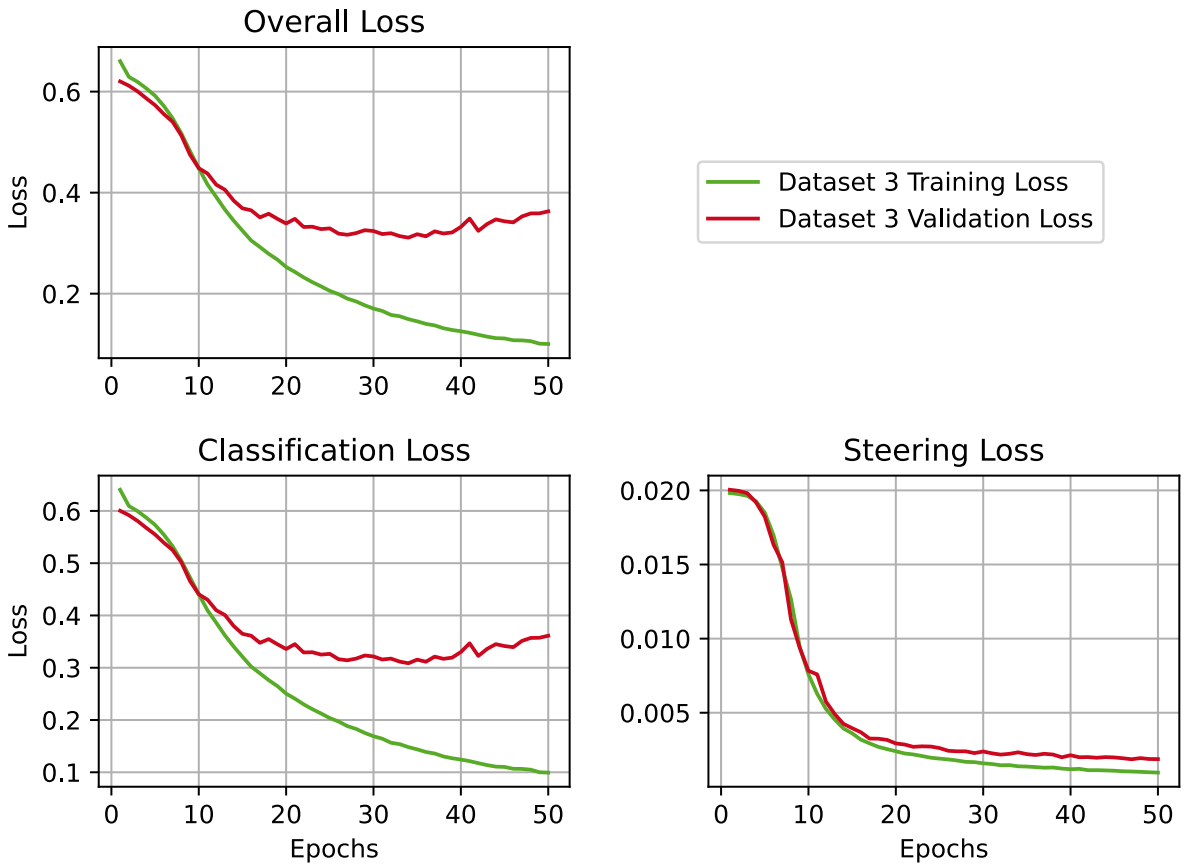


Figure 5.13: Separate LSTM layers for classification and steering

Criteria(Softmax/Categorical crossentropy). Rating 1 to 5. 1 being the lowest	Dataset 3
Lane keeping/Drive straight	4
Gradual acceleration increase	3
Smooth braking behaviour observed	3
Smooth steering control at high speed(10m/s)	4
Smooth steering control at turnings	1
Avoids colliding with static objects	2
Detects vehicles as dynamic objects	5
Navigates traffic smoothly	4
Stops at random places	5
Smooth evaluation experience	3

Table 5.9: Split at the LSTM layer - How the model evaluates to different criteria

traffic, the model does well to navigate, brake, and accelerate.

A quick overview of steering losses across different NN changes is shown in fig. 5.16.

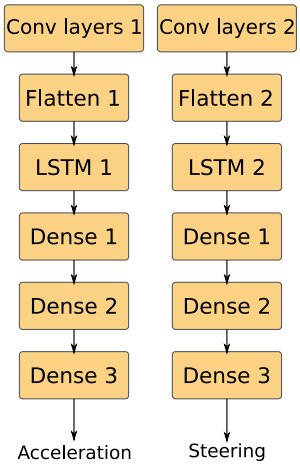


Figure 5.14: Separate NN training model

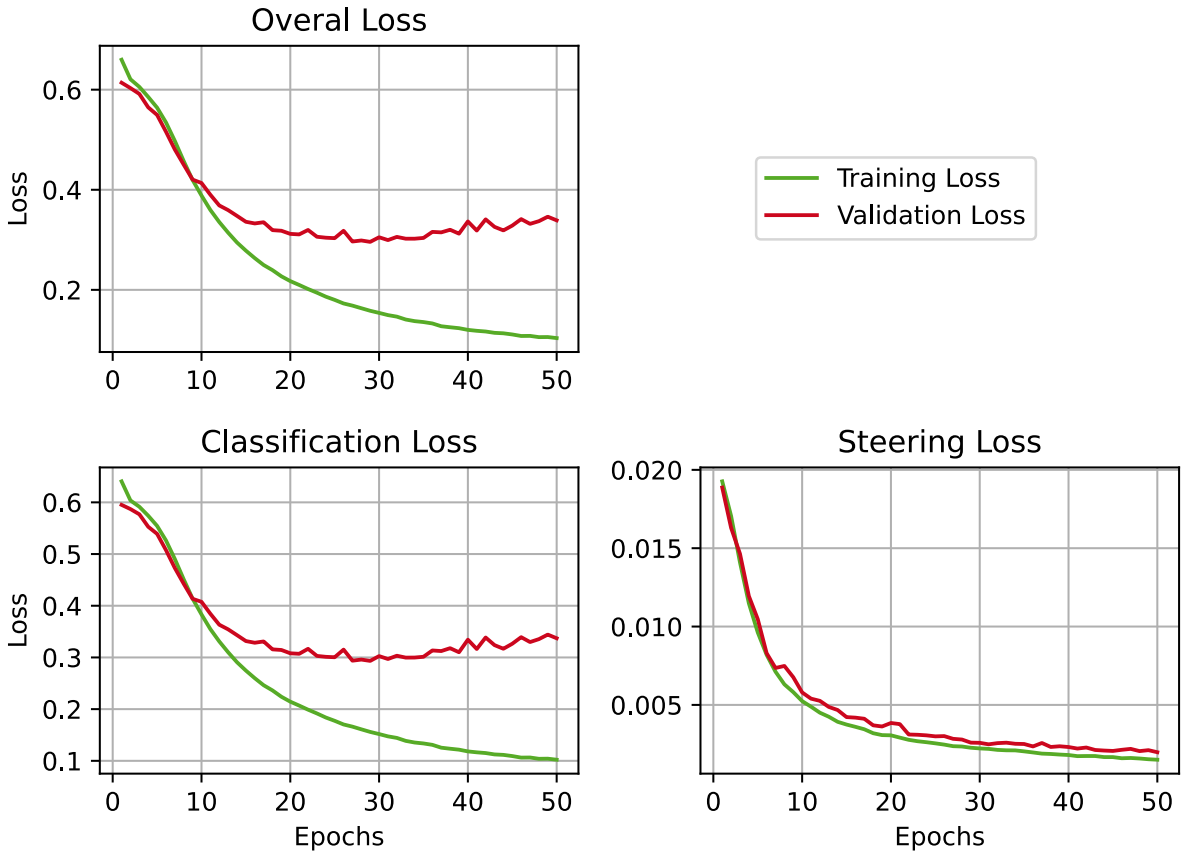


Figure 5.15: Separate neural network for Classification and Steering

5.3.5 Observations

1. Tuning the neural network albeit only the fully connected layers, results in optimised prediction of steering control corresponding to the classification outputs.

2. The car exhibits random, unknown stops at random places. The actual reason behind it is unknown but it is suspected that since the dataset has imbalanced classes, it contributes to this random decisions.

Criteria(Softmax/Categorical crossentropy). Rating 1 to 5. 1 being the lowest	Dataset 3
Lane keeping/Drive straight	4
Gradual acceleration increase	4
Smooth braking behaviour observed	4
Smooth steering control at high speed(10m/s)	4
Smooth steering control at turnings	2
Avoids colliding with static objects	3
Detects vehicles as dynamic objects	5
Navigates traffic smoothly	3
Stops at random places	5
Smooth evaluation experience	4

Table 5.10: Separate neural network for classification and steering outputs - How the model evaluates to different criteria

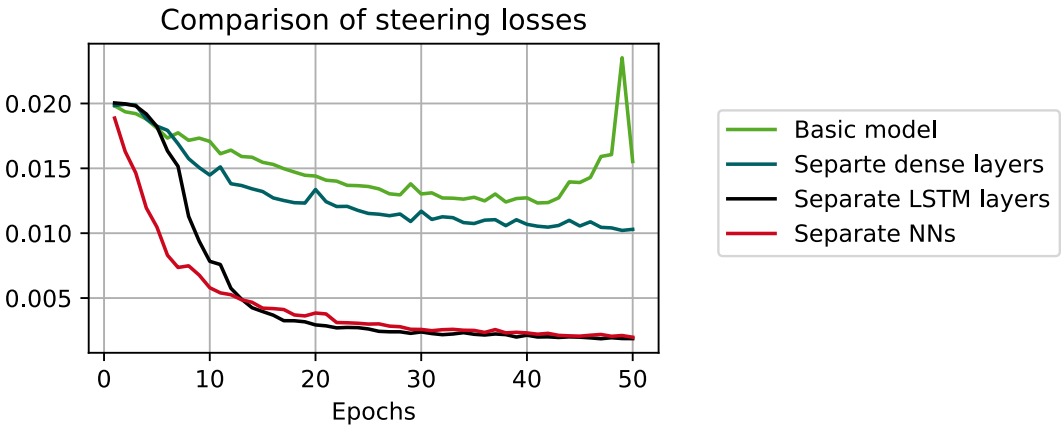


Figure 5.16: Steering command loss comparison

3. Separating into two NNs allows better steering control at a higher speed.

4. Sunlight and shadows still play a major role. They do some random, strange things to models that eventually lead to crashes out-of-nowhere. It could be deduced that some buildings' shadows could be considered as static or dynamic objects.

5.4 Velocity

Velocity is a scalar value, labelled output. It is considered as an *auxiliary task*. An auxiliary task usually consists of estimating quantities that are relevant to solving main supervised learning problem. That means that velocity is predicted as an auxiliary task using existing CNN-LSTM-Dense architectures without affecting the major tasks i.e., predicting acceleration and steering.

The images are fed into the models shown in fig. 5.17 and results are compared as how predicting velocity affects acceleration and steering.

From the fig. 5.18, looking at the loss-epoch graphs of model *a*, *b*, *c*, the validation losses follow velocity's validation loss. When compared to classification and steering loss, this loss is too high. Sure enough while evaluating steering and acceleration are worse. Since velocity is an auxiliary task, there is a need to rethink how cost function is calculated.

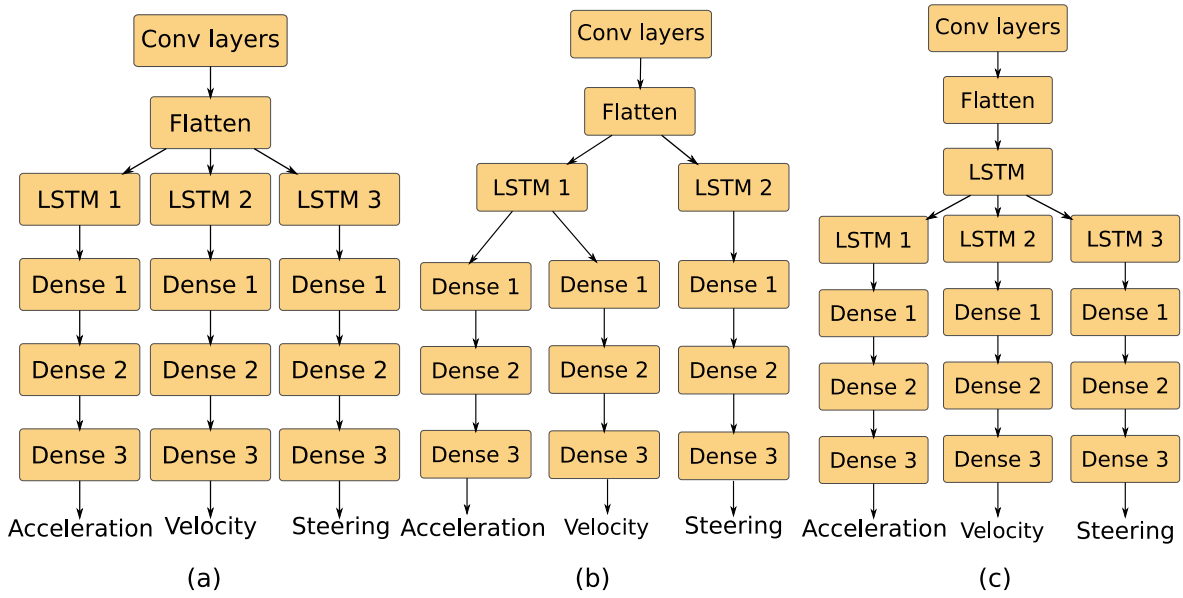


Figure 5.17: Different architectures used while predicting velocity

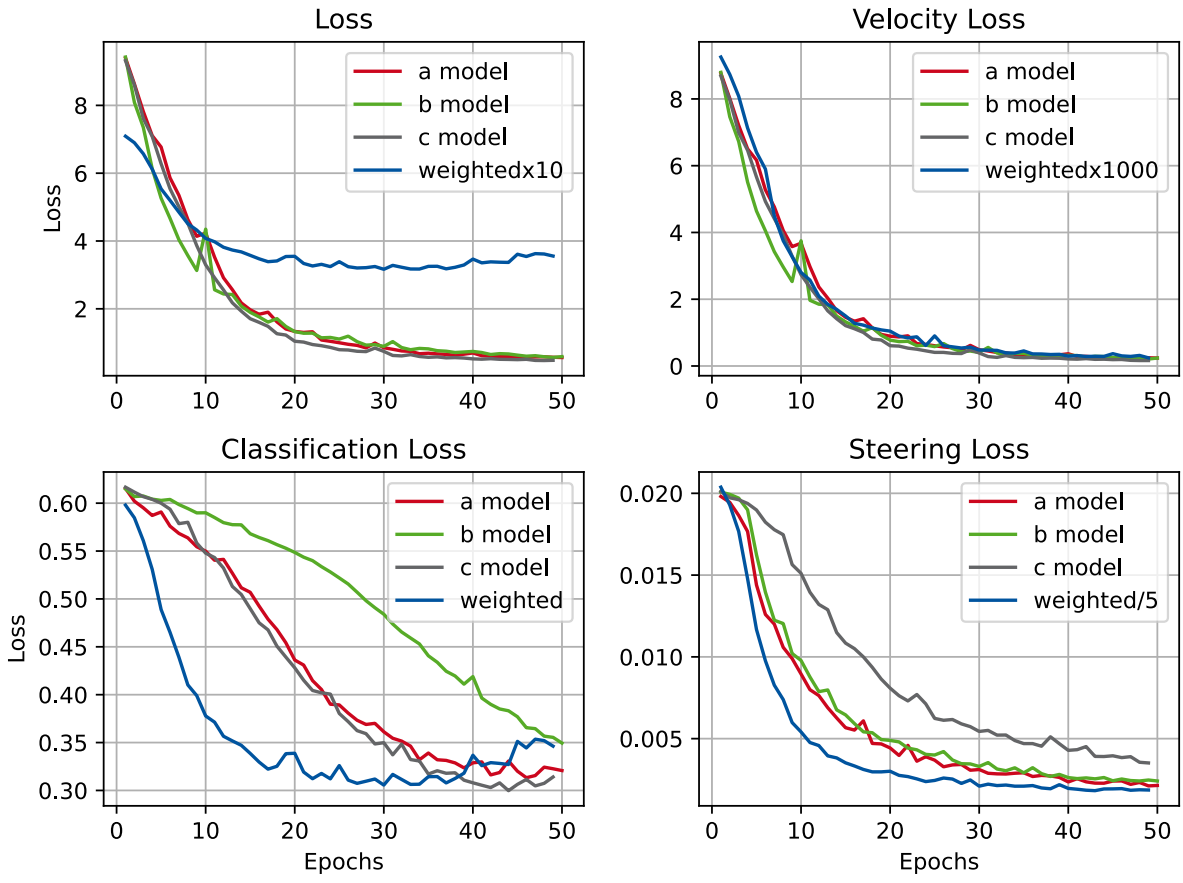


Figure 5.18: Comparison of losses for NN architectures shown in fig. 5.17

5.4.1 Weighted loss function

Since velocity is taking control of the neural network to provide feature extraction and decision making for its prediction, both acceleration and steering suffer bad consequences. In order to avoid this, weighted cost function is introduced to the three outputs. As velocity is an auxiliary task, it is suppressed more than others. From figures 5.17 and 5.18, model *a* gives nearly optimum losses. So that model is chosen to carry out this weighted cost function experiment.

The model is tweaked only to include custom, weighted cost function. The trained model exhibits now losses similar to classification losses than velocity(as seen in fig. 5.18, *weighted* graph). Indeed during evaluation, acceleration and steering both get preference and control than velocity. Hence it is possible to include auxiliary tasks to normal NNs without compromising the actual functionality of it.

Observation

1. Including an auxiliary task such as velocity needs to be custom weighted to allow the primary function of the NN to be fulfilled.

2. Velocity prediction has a high loss. More analysis needs to be done tweaking the NN model differently and running experiments.

5.5 Convolution layers manipulation

The fully connected/dense layers are tweaked to various designs as shown in figures 5.8, 5.10, 5.12, 5.14 and having convolutional layers as constant. Some interesting possibilities and observations are made possible. As a consequence, the convolutional layers are changed keeping dense layers as constant. As velocity prediction is also included, 5.17a model is considered.

The convolutional layers can be adjusted either by width(changing the feature maps or stride parameter) or depth(changing the number of layers).

5.5.1 Adjusting the width of the convolutional layers

Changing the feature maps channel depth

The convolutional layers consists of feature maps channels, kernel filter which convolves on the input using a specified stride. In our case, the last convolutional layer's feature map channel's depth is increased from 64 to 80 (as shown in fig. 5.19a). This change, after *flatten* layer, increases the trainable parameters. This allows for more features to carried into the fully connected layer. Looking at the loss-epoch graph(fig 5.20) for this experiment, they give relatively similar results as 5.18a model.

Changing the stride

Stride is a component in CNN tuned for compressing the images data. This parameter specifies the movement step of the kernel/filter. In our case, the stride of the 5th convolutional layer is decreased from (2,2) to (1,1) (see fig. 5.19b). This ensures the information tensors are not compressed by half. The flatten layer gets uncompressed, more tensors which increases the trainable parameters of the features. From fig. 5.20, we see that losses are similar to the 5.18a model but the classification loss starts to overfit well ahead of the *a* model. Since the training model follows the classification loss, the best model is not stored with this change.

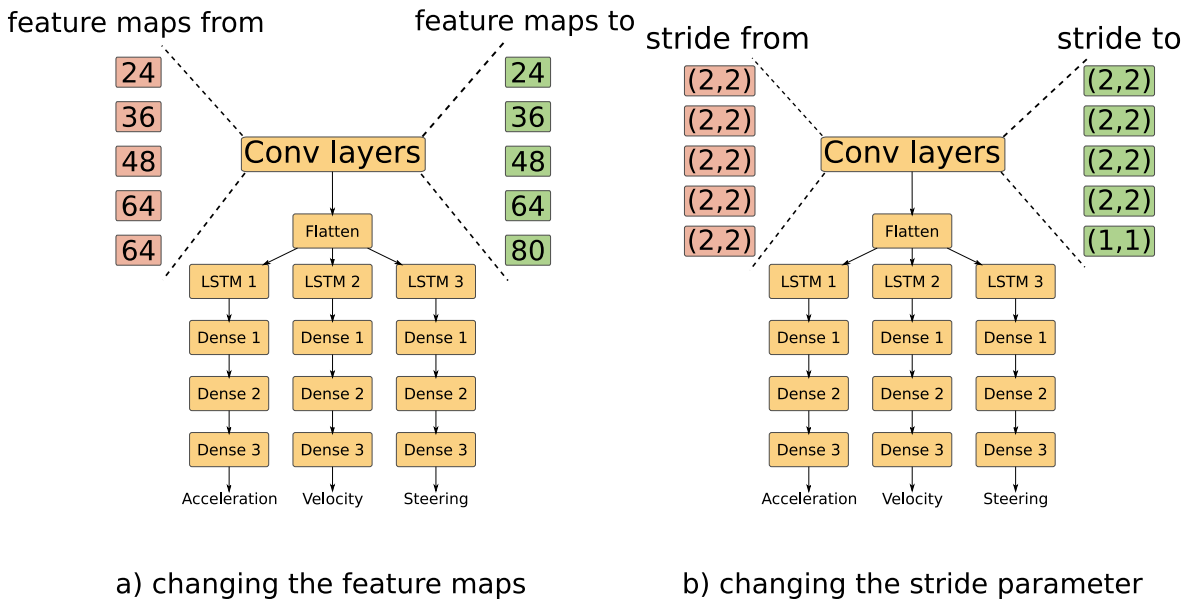


Figure 5.19: Convolutional layers width changes - Increasing feature maps channel depth

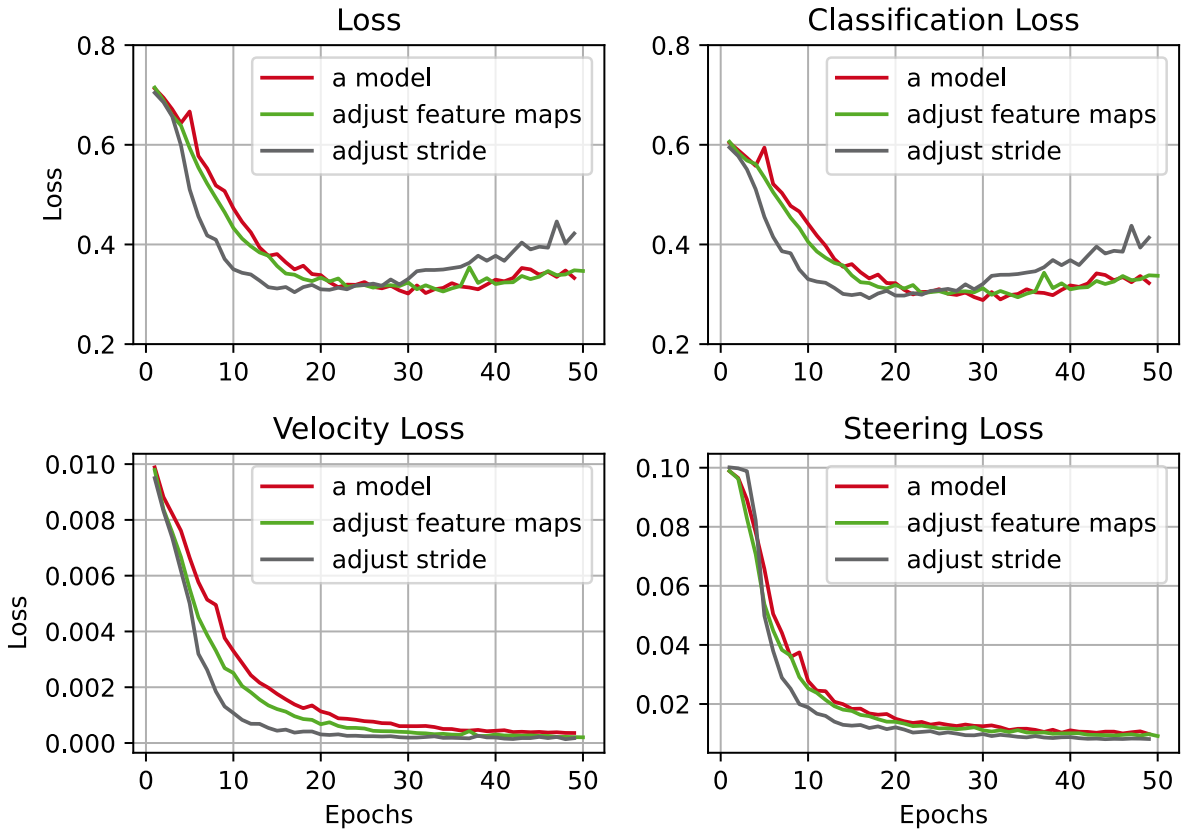
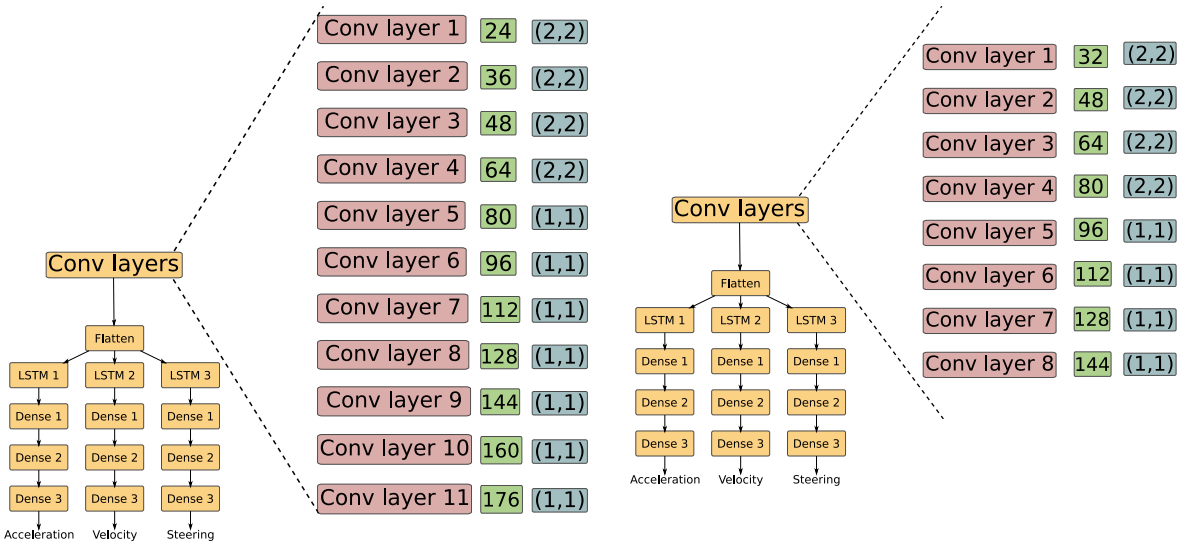


Figure 5.20: Convolutional layers width adjustments - Comparison of losses

5.5.2 Adjusting the depth of the Convolutional layers

Increase the convolutional layers to eleven

Fig. 5.21a, shows how the convolutional layers are increased to 11 layers and also the tensor shape before *flatten* layer is made greater than the previous case(a model). The change in losses because of these adjustments is instant. Fig. 5.22 shows that the model overfits as soon as epoch 10. The model follows the classification loss. Though other losses continue to decrease, these events are not recorded.



a) Increasing to 11 layers b) Increasing to 8 layers

Figure 5.21: Convolutional layers depth changes - Increasing the number of CNN layers

Increase the convolutional layers to eight

If the convolutional layers are increased only to 8 and the tensor shape is kept as small as possible before flatten layer, the loss-epoch graph 5.22 shows better behaviour than increasing to 11 layers. However, this change also brings about overfitting behaviour. It could be seen that the velocity and steering losses converge well. However, like before, they are not recorded.

5.6 Depth Camera

The depth camera is typically used to measure the distance between objects. In our case, we use it to provide additional spatial information between vehicles and objects. We test how the neural networks adapts and predicts acceleration, braking and steering.

The fig. 5.23 loss-epoch graphs show a little better performance compared to colour-RGB images trained model. For training the architecture 5.17a is used. Upon evaluation, however, as populated in the table 5.11, the results are evident; it is bad. Because the classification loss is bad, the acceleration doesn't work at all. Steering works but since depth images don't have features such as lane markings in their images, the car can't right itself when it is on the boundary.

Surprisingly, the model recognises traffic vehicle when driven at a very slow speed. It even stops in

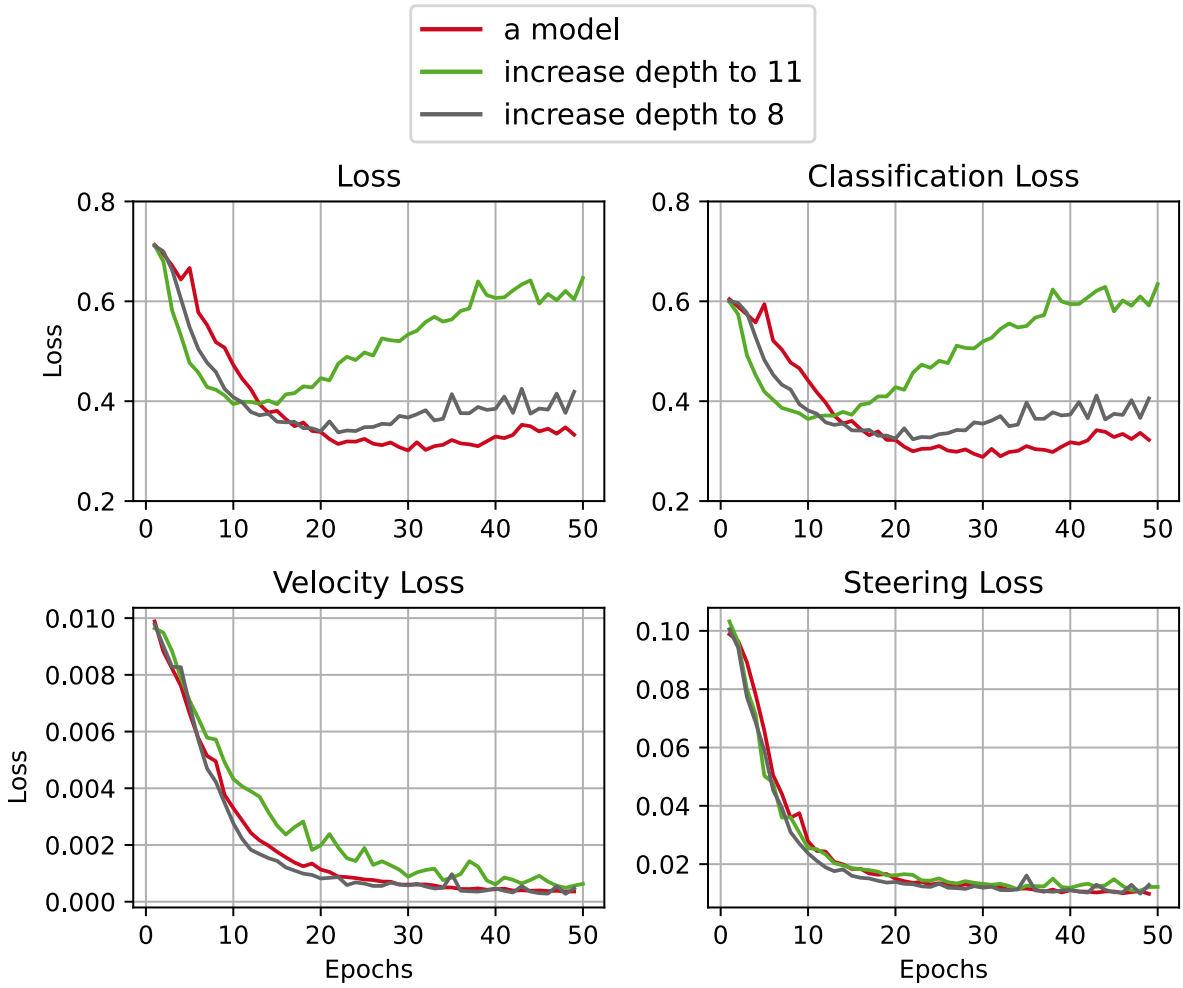


Figure 5.22: Convolutional layers depth adjustments - Comparison of losses

front of traffic in some situations. Static objects are also recognised and sometimes the model predicts correct steering angles to avoid collision.

So this standalone training lets us to observe that with depth sensor images, it is possible to recognise other vehicles and static objects like pavement. In the next section, we explore data fusion by combining RGB and depth images.

5.7 Segmentation camera images

The segmented images consist of instance tags which are initialised in the LGSVL sensor parameters. They provide colour instance of each object on the map. For eg. car is pink, pedestrian red, barriers white etc. So a training with just segmented camera images is done and the losses-epoch graph(fig. 5.24) is plotted. The graph shows that it performs similar to RGB-Grayscale training model.

5.8 Data Fusion

It is evident that using a different source of image other than RGB, makes the model not recognise enough features. In RGB image models alone we see the velocity prediction is bad. So we need

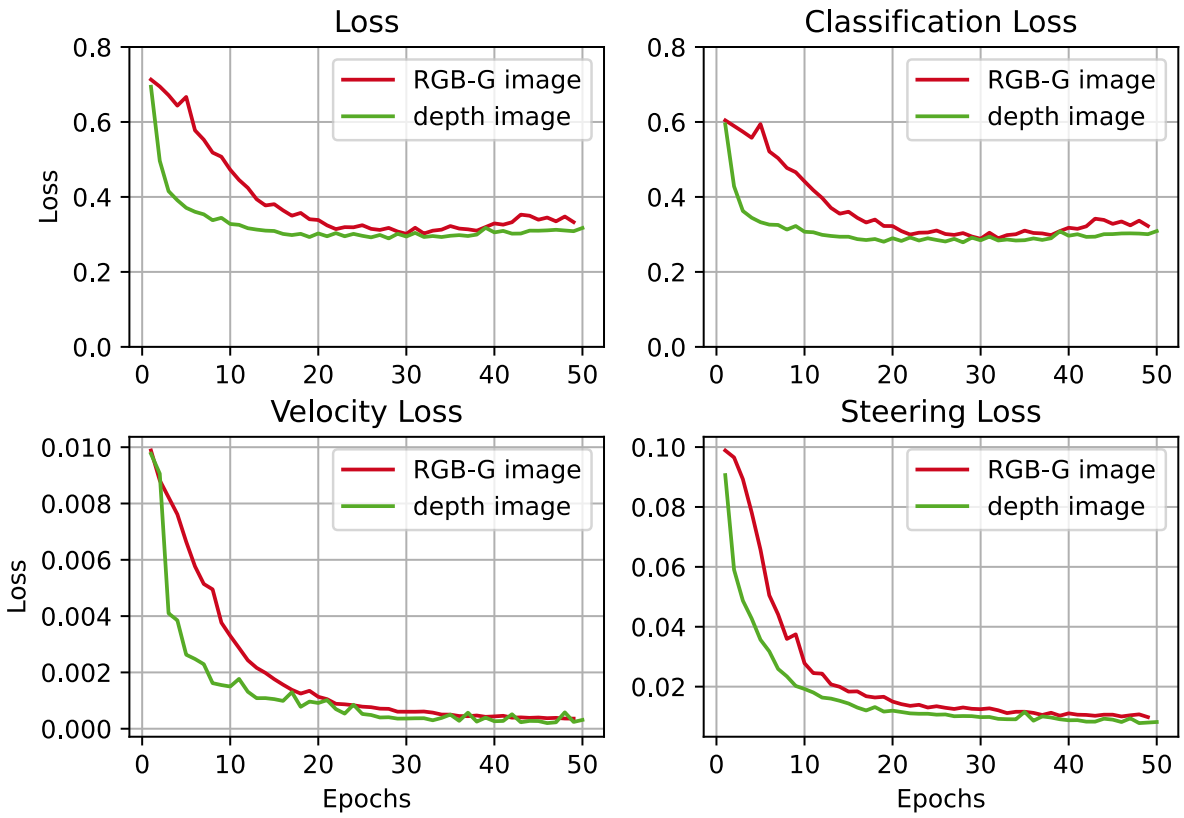


Figure 5.23: Comparison of losses between using RGB-Grayscale and depth images

Criteria. Rating 1 to 5. 1 being the lowest	Rating
Lane keeping/Drive straight	1
Gradual acceleration increase	1
Smooth braking behaviour observed	1
Smooth steering control at high speed(10m/s)	1
Smooth steering control at turnings	1
Avoids colliding with static objects	3(at slow speed)
Detects vehicles as dynamic objects	4
Navigates traffic smoothly	1
Stops at random places(negative case)	5
Smooth evaluation experience	1

Table 5.11: Depth sensor evaluation - standalone

something more than RGB images to predict these parameters.

5.8.1 Early fusion

As explained in chapter 2, section 2.4, early fusion involves combining different sources as one and feeding into the network. Here we take different sources of images such as RGB, depth, and segmented images, fuse two or all of them together and feed it to the NN as shown in fig. 2.16a.

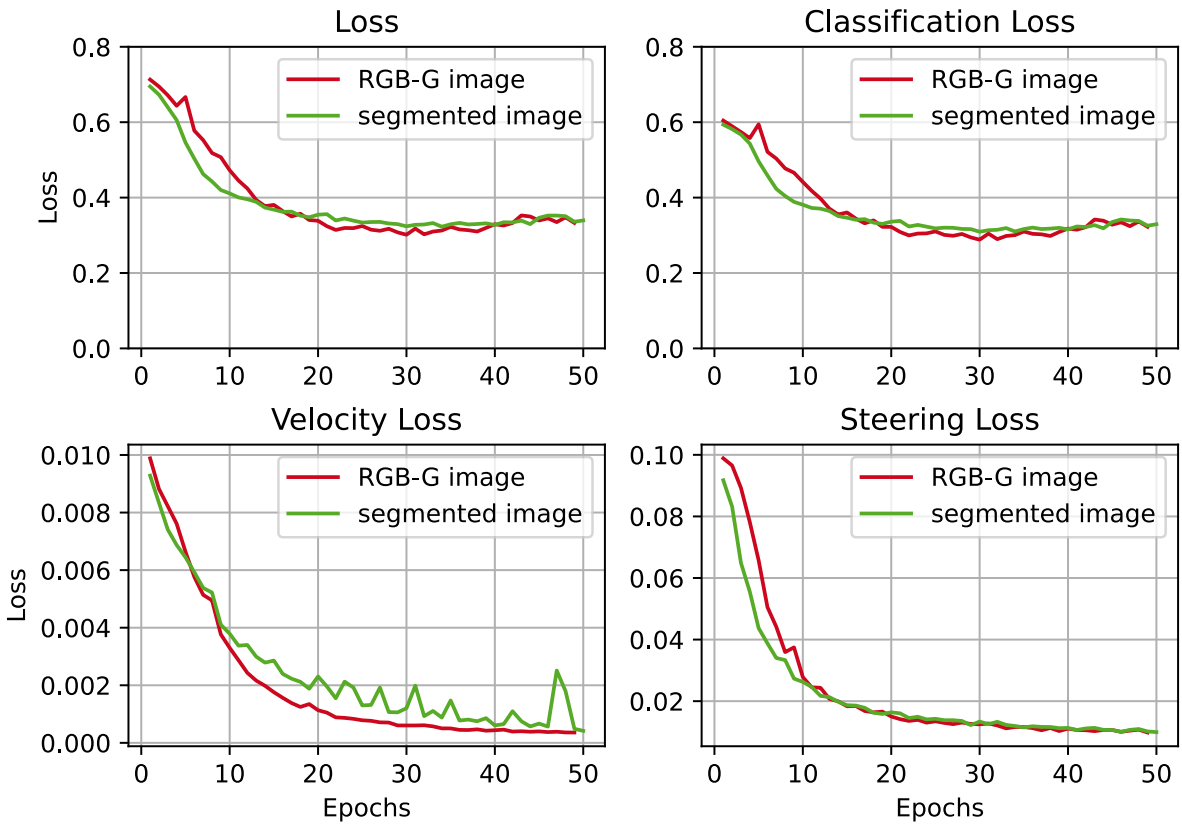


Figure 5.24: Comparison of losses between using RGB-Grayscale and Segmented images

RGB-Grayscale and Depth images

The table 5.12 gives the changes in hyperparameters made while conducting this technique.

Setting	Value
Feature maps	[24, 36, 48, 64, 80]
Input	(15, 70, 160, 2)

Table 5.12: Hyperparameter setting changes - Early fusion

From the graph 5.25, it can be seen that early fusion of RGB-Grayscale and depth images, performs better than standalone versions of either of them. The overall loss which follows the classification loss is lower than standalone variants. However, looking at loss graph alone isn't a good judge of the model.

Sure enough, upon evaluation, table 5.13 shows the model performs slightly better to static objects. However, it fails to acceleration consistently and braking is observed regularly observed.

RGB-Grayscale and Segmented images

For these image sources, the same hyperparameter setting used for RGB-G-Depth is also used here.

From the graph 5.25, this combination of images performs almost similar to standalone RGB-G. It might be because the segmented has only 11 instances of objects where they are coloured differently using tags. This image, however, due to limitation of computing resources, is converted to grayscale,

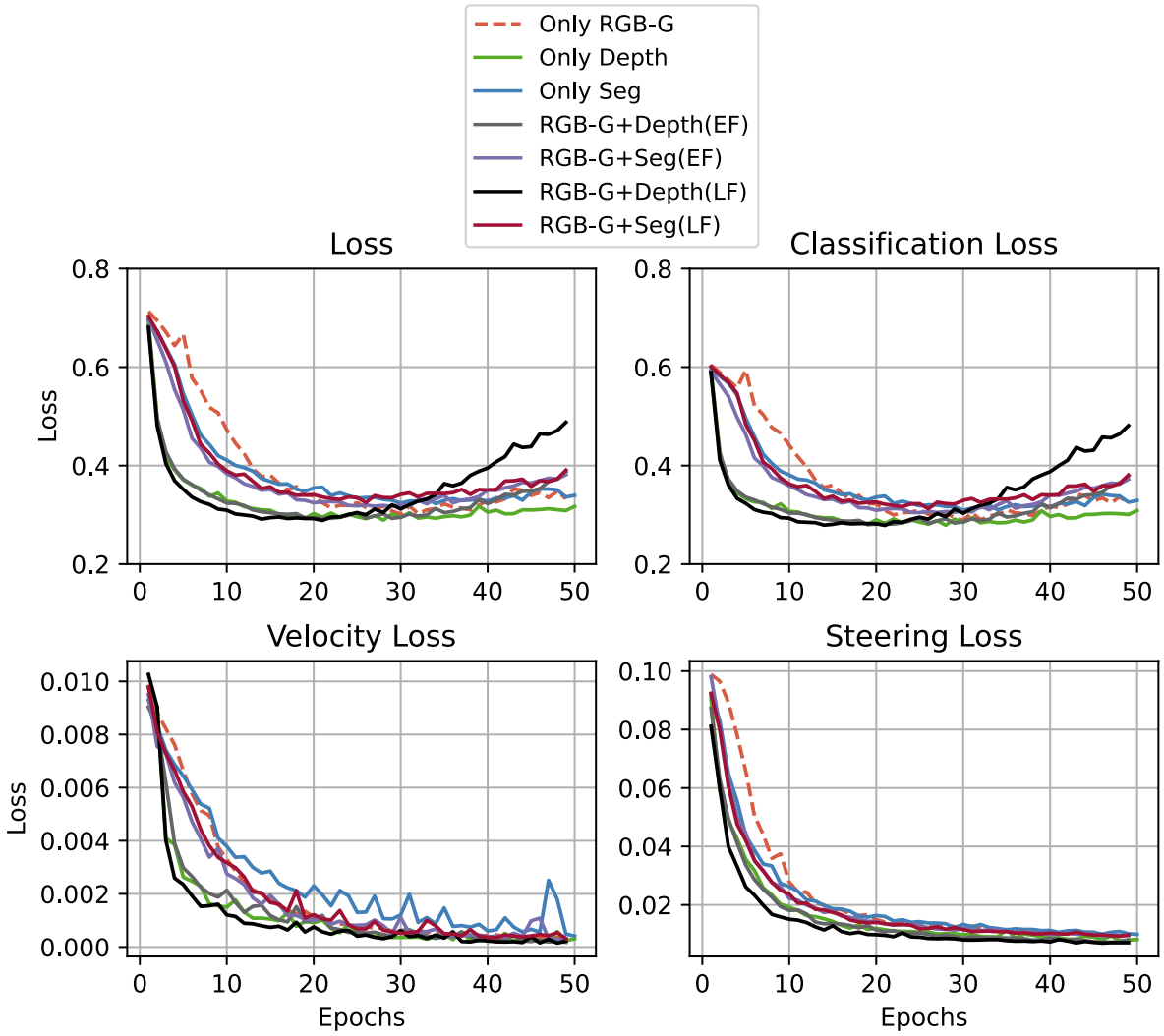


Figure 5.25: Comparison of fusion losses

Criteria. Rating 1 to 5. 1 being the lowest	Rating
Lane keeping/Drive straight	3
Gradual acceleration increase	3
Smooth braking behaviour observed	1
Smooth steering control at high speed(10m/s)	1
Smooth steering control at turnings	1
Avoids colliding with static objects	3(at slow speed)
Detects vehicles as dynamic objects	4
Navigates traffic smoothly	1
Stops at random places	5
Smooth evaluation experience	1

Table 5.13: Early fusion evaluation - RGB-Grayscale and Depth images

taking away some of its ability.

After evaluation table 5.14 is tabulated with results. The results are positive and the car navigates smoothly when introduced to traffic.

Criteria. Rating 1 to 5. 1 being the lowest	Rating
Lane keeping/Drive straight	4
Gradual acceleration increase	4
Smooth braking behaviour observed	2
Smooth steering control at high speed(10m/s)	3
Smooth steering control at turnings	4
Avoids colliding with static objects	3
Detects vehicles as dynamic objects	5
Navigates traffic smoothly	2
Stops at random places	5
Smooth evaluation experience	4

Table 5.14: Early fusion evaluation - RGB-Grayscale and Segmented images

5.8.2 Late fusion

Similar to early fusion, late fusion explained in section 2.4, is attempted using 2.16b as reference. The images from different sources are separately fed to the NN and after two convolutional layers, their outputs are fused. Then the fused output is continued with the remaining layers.

Setting	Value
Feature maps	[24, 36] [24, 36], [80, 96, 96]
Input	(15, 70, 160, 2)

Table 5.15: Hyperparameter setting changes - Late fusion

RGB-Grayscale and Depth images

The table 5.15 gives the changes in hyperparameters made while conducting this technique.

We have seen from loss graph 5.22, increasing the capacity of convolution layers for these images' dimension, makes the model overfit quicker than standalone variants. Just so, the graph for this experiment 5.25, overfits earlier and faster than other methods. We, however, see a minuscule gain in classification loss and good improvement in velocity and steering losses.

To verify these observation, the model is tested and table 5.16 is filled up. Just like its early fusion counterpart, the prediction is not good as hoped from the losses graph. It does certain things really well but not on consistent basis.

RGB-Grayscale and Segmented images

The losses 5.25 for this experiments are nearly identical to its early fusion variants. Indeed while evaluating, the model performs great to tough situations such as avoiding collision to barriers or pavements. Of course, there are a few problems which need further analysis to resolve.

Criteria. Rating 1 to 5. 1 being the lowest	Rating
Lane keeping/Drive straight	4
Gradual acceleration increase	1
Smooth braking behaviour observed	1
Smooth steering control at high speed(10m/s)	1
Smooth steering control at turnings	1
Avoids colliding with static objects	3
Detects vehicles as dynamic objects	3
Navigates traffic smoothly	1
Stops at random places	5
Smooth evaluation experience	1

Table 5.16: Late fusion evaluation - RGB-Grayscale and Depth images

Criteria. Rating 1 to 5. 1 being the lowest	Rating
Lane keeping/Drive straight	5
Gradual acceleration increase	2
Smooth braking behaviour observed	3
Smooth steering control at high speed(10m/s)	3
Smooth steering control at turnings	4
Avoids colliding with static objects	4
Detects vehicles as dynamic objects	5
Navigates traffic smoothly	3
Stops at random places	5
Smooth evaluation experience	4

Table 5.17: Late fusion evaluation - RGB-Grayscale and Segmented images

Observations

1. Fusing different sources give new dimension to the model which helps in making more correct decisions.

2. Fusing RGB-G and depth images don't seem to match the expectations.

3. Fusing RGB-G and segmented images produce some great results. It would be a good avenue to pursue future analysis.

5.9 Extending the RGB-G+Segmented early fusion to larger dataset

With the promise shown by RGB-G and segmented images fusion evaluation, it is necessary to try to increase the dataset with more robust data, train the model, and test it. Hence the dataset 3 is increased from 100,000 to 224,000 data entries consisting also change in weather conditions and night time driving. With this dataset, the same setup as RGB-G and segmented image, early fusion is used. From the losses-epoch graph fig. 5.26, the classification losses can be prevented from overfitting by increasing the data and manipulating the learning rates. For this training alone, learning rate is started

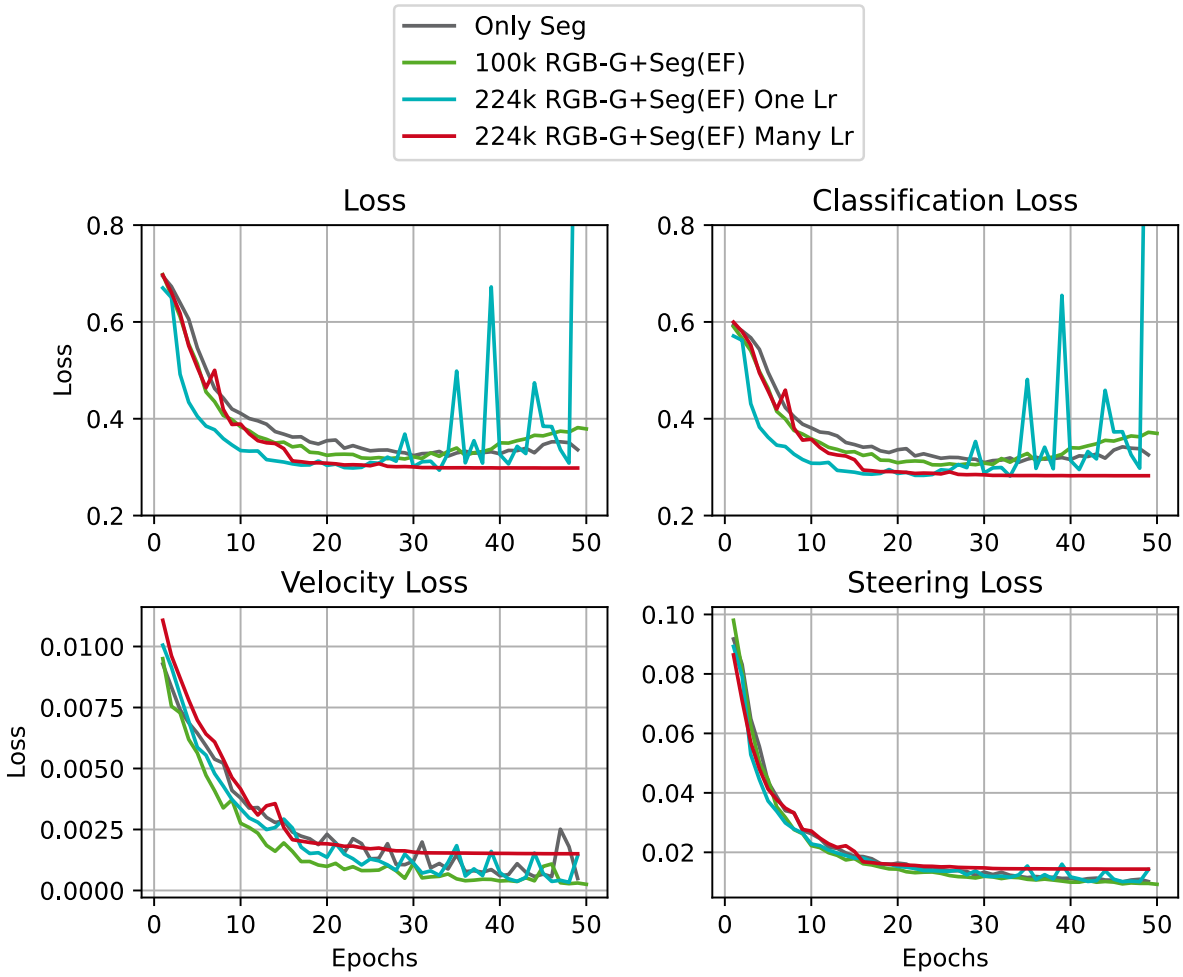


Figure 5.26: Comparison of losses between Segmented images vs 100k RGB-G+Seg vs 224k RGB-G+Seg models. *Lr* denoted learning rate.

with 10^{-4} and continued to decrease every 20 epochs. Since the learning rate is decreased, the training epochs is increased to 150-200. However, for the sake of comparison only 50 epochs is shown.

The table 5.18 shows training with more data helps in performance. The model is unaffected by sunlight, weather changes or time of the day. It performs quite good even in night time conditions. This is possible only because of segmentation sensor camera images as it clearly defines the lane boundaries, cars etc., Though, there is less-to-no control at high speeds, this training model performs the best among all other trainings.

Observations

1. Adding more data to the training helps in reducing the classification loss further resulting in a little better prediction of acceleration.
2. With the help of new labelled data exposed different weather conditions, the trained model now performs good at wet conditions.
3. Night time driving which was impossible(see fig. 5.2a), is now made possible with fusing segmentation sensor images. The chart 5.27 shows good improvement in number of collisions while driving at night.

Criteria. Rating 1 to 5. 1 being the lowest	Rating
Lane keeping/Drive straight	5
Gradual acceleration increase	3
Smooth braking behaviour observed	2
Smooth steering control at high speed(10m/s)	2
Smooth steering control at turnings	**5**
Avoids colliding with static objects	**5**
Detects vehicles as dynamic objects	4
Navigates traffic smoothly	2
Affected by sunlight	**4**
Affected by low-light or night driving	**4**
Stops at random places	5
Smooth evaluation experience	5

Table 5.18: Larger dataset early fusion evaluation - RGB-Grayscale and Segmented images

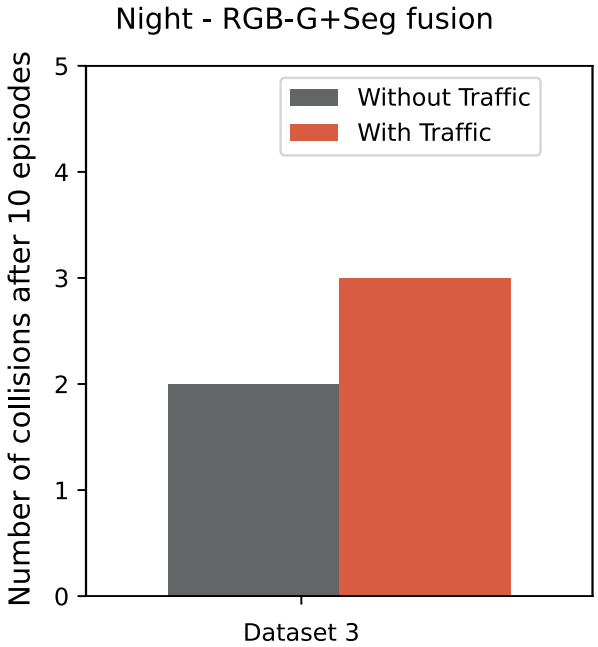


Figure 5.27: Night time driving with RGB-G and Segmented images fusion - Traffic vs collisions

6 Conclusion

In this thesis, an end-to-end neural network for autonomous driving with basic framework The ROS 2 is a flexible framework which enables fast communication between components. The version of ros-bridge used for this project, ros2-web-bridge, carries too much overhead and affects the performance. It is advised to use the new lgsvl-ros2-bridge from the LGSVL team instead. LGSVL simulator provides a great solution to collect and test algorithms. In about 24 hours, it is possible to collect more than 250,000 data entries and label them. is implemented. Upon evaluating, we observe the following:

1. Light plays a major role on how good the model predict the correct decision as shown in 5.1.
2. Steering performs remarkably well even when trained just with RGB-Grayscale image.
3. Acceleration and braking work best when performed as classification tasks.
4. First dataset (dataset 1) though was collected without traffic, performs excellently when exposed to it. It also deals with dynamic and static objects which signifies the generalisation example of supervised learing.
5. All the datasets(5.6) are imbalanced with *no action* state taking majority of the entries. This causes a bias in decision making.
6. Splitting the fully connected layers help in optimising the weights calculations resulting in somewhat balanced prediction.
7. Including a highly varying auxiliary task such as velocity skews the model to predict it than its primary task. Therefore weighting the cost function is most useful in suppressing the velocity and enhancing the primary task.
8. Adjusting the width and depth of the convolutional layers is almost as difficult as adjusting the fully connected layer. Increase in trainable parameter doesn't really mean more features.
9. Both depth and segmented images carry no feature information such as lane markings which make it harder to predict outputs using only them. They are, however, highly useful when combined with a RGB-Grayscale image.
10. Late fusion takes longer than early fusion technique as the fusing operation is done while training the model.
11. RGB-Grayscale plus segmented images fusion give much better results and perform really well even at night and wet conditions.
12. While evaluation, bigger and more complex the prediction NN model, longer it takes to predict. This will then delay the output being published to the simulator.

To conclude, this project with the help of docker, ROS, LGSVL, Keras, Tensorflow, and Python, it is possible to construct a basic end-to-end model that incorporates state of the art data fusion technique to help make better decisions. However, without regular feedback by testing the algorithm from real-world conditions, significant progress in this field would be difficult.

7 Future Work

One of the motivations of this thesis is to see whether it is possible to use LGSVL simulator to implement an end-to-end network to predict control commands by using supervised learning. The results prove that the controlling the car using predicted commands just by using images is possible.

There are, however, several drawbacks which in future can be improved.

1. Make the dataset more balanced and robust so that the classifier can make unbiased decisions and also make the trained model not affected by sunlight.

2. Dataset 1 did well in traffic even though it was not exposed to one during data collection phase. Dataset 3 on the other hand, failed miserably in evaluation when performed as a regression task. A combination of these two datasets made it possible to achieve better results. So a dataset consisting of holding the lane and careful driving would help in achieving overall better performance.

3. Though dataset 3 had about 270,000 data entries, the computer didn't have the necessary computing power or the resources to hold such amount of data. More computational power would help make faster, bigger, and robust models. Almost all of the training were carried out with input image's dimension as (160,70,1). A significant increase in computing resources will help in increasing image dimensions and convolutional layers for better feature extraction.

4. Auxiliary task as predicting velocity along with acceleration and steering overwhelmed the model if not weighted properly. Still velocity prediction was worse. Some work could be done to include these tasks efficiently.

5. Data fusion with depth images didn't function as expected. Future analysis as to why this is happening and how to rectify it would be a good area of research.

6. By increasing the data entries to 224,000, it is shown the classification loss can be prevented from overfitting. So more, balanced data will give more favourable results.

List of Figures

List of Tables

References

[1] Google, “Waymo.” https://waymo.com/

[2] Berkeley University, “Berkeley-deepdrive. [online].” https://deepdrive.berkeley.edu/

[3] Apollo, “Apolloscape dataset.” http://apolloscape.auto/

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[5] H. Rohling and E. Lissel, “77 ghz radar sensor for car application,” in *Proceedings International Radar Conference*, 1995, pp. 373–379.

[6] D. Lv, X. Ying, Y. Cui, J. Song, K. Qian, and M. Li, “Research on the technology of lidar data processing,” in *2017 First International Conference on Electronics Instrumentation Information Systems (EIIS)*, 2017, pp. 1–5.

[7] A. Carullo and M. Parvis, “An ultrasonic sensor for distance measurement in automotive applications,” *IEEE Sensors Journal*, vol. 1, no. 2, pp. 143–, 2001.

[8] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta *et al.*, “Lgsvl simulator: A high fidelity simulator for autonomous driving,” *arXiv preprint arXiv:2005.03778*, 2020. https://www.lgsvlsimulator.com/

[9] Nvidia, “Nvidia simulator.” https://www.nvidia.com/en-us/self-driving-cars/drive-constellation/

[10] Carla, “Carla simulator.” https://carla.org/

[11] IPG-Automotive, “Carmaker- simulator.” https://ipg-automotive.com/products-services/simulation-software/carmaker/

[12] Open Robotics, “Robotic operating system.” https://index.ros.org/doc/ros2/

[13] Seth Lambert and Erika Granath, “Lidar systems: costs, integration, and major manufacturers.” https://www.mes-insights.com/lidar-systems-costs-integration-and-major-manufacturers-a-908358/

[14] Z. Chen and X. Huang, “End-to-end learning for lane keeping of self-driving cars,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1856–1860.

[15] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, “End to end learning for self-driving cars,” 2016.

[16] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *The International Journal of Robotics Research*, 03 2016.

[17] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," *CoRR*, vol. abs/1612.01079, 2016. http://arxiv.org/abs/1612.01079

[18] J. Kim and J. F. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," *CoRR*, vol. abs/1703.10631, 2017. http://arxiv.org/abs/1703.10631

[19] M. Bojarski, P. Yeres, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, and U. Muller, "Explaining how a deep neural network trained with end-to-end learning steers a car," *CoRR*, vol. abs/1704.07911, 2017. http://arxiv.org/abs/1704.07911

[20] Z. Yang, Y. Zhang, J. Yu, J. Cai, and J. Luo, "End-to-end multi-modal multi-task vehicle control for self-driving cars with visual perceptions," pp. 2289–2294, 08 2018.

[21] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning," 2018.

[22] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition." in *WACV*. IEEE Computer Society, 2016, pp. 1–8. http://dblp.uni-trier.de/db/conf/wacv/wacv2016.html#ParkHBB16

[23] Y. Zhou and K. Hauser, "Incorporating side-channel information into convolutional neural networks for robotic tasks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2177–2183.

[24] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 695–12 705.

[25] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *CoRR*, vol. abs/1906.03199, 2019. http://arxiv.org/abs/1906.03199

[26] F. Codevilla, M. Mller, A. Lopez, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4693–4700.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[28] T. Mitchell, "Machine learning," *McCraw Hill*, 1996.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, p. 100, http://www.deeplearningbook.org.

[30] K. P. Murphy, *Machine learning: a probabilistic perspective*, Cambridge, MA, 2012, p. 3.

[31] Canadian Institute for Advanced Research, "Cifar 10 image dataset." https://www.cs.toronto.edu/~kriz/cifar.html

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[33] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 21212159, jul 2011.

[34] G. Hilton, "RMSProp - Unpublished method from coursera lecture," 2012. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

[35] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, p. 85117, Jan 2015. http://dx.doi.org/10.1016/j.neunet.2014.09.003

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 06 2014.

[37] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[38] S. Hochheiter and J. Schmidhuber, "Untersuchungen zu dynamischen neuronalen netzen. masters thesis, technische universitaet muenchen," *Diplomarbeit*, June 1991. http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[40] B. Cain, "Passive Stereo Camera." https://brenn10.github.io/MIT_experience_paper/index.html

[41] U. Hahne and M. Alexa, "Combining time-of-flight depth and stereo images without accurate extrinsic calibration," *Int. J. Intell. Syst. Technol. Appl.*, vol. 5, no. 3/4, p. 325333, nov 2008. https://doi.org/10.1504/IJISTA.2008.021295

[42] A. Lipnickas and A. Kny, "A stereovision system for 3-d perception," *Elektronika ir Elektrotechnika*, pp. 1392–1215, 04 2009.

[43] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *CoRR*, vol. abs/1902.04502, 2019. http://arxiv.org/abs/1902.04502

[44] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, p. 2844, jan 2013. https://doi.org/10.1016/j.inffus.2011.08.001

[45] V. Malyavej, W. Kumkeaw, and M. Aorpimai, "Indoor robot localization by rssi/imu sensor fusion," in *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2013, pp. 1–6.

[46] Y. Dobrev, S. Flores, and M. Vossiek, "Multi-modal sensor fusion for indoor mobile robot pose estimation," in *2016 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, 2016, pp. 553–556.

[47] D. Iter, J. Kuck, and P. Zhuang, "Target tracking with kalman filtering, knn and lstms," *proj2016*, 2016. http://cs229.stanford.edu/proj2016/report/IterKuckZhuang-TargetTrackingwithKalmanFilteringKNNandLSTMs-report.pdf

[48] H. Coskun, F. Achilles, R. DiPietro, N. Navab, and F. Tombari, "Long short-term memory kalman filters:recurrent neural estimators for pose regularization," 2017.

[49] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, "Spatially supervised recurrent convolutional neural networks for visual object tracking," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4.

[50] S. Bohez, T. Verbelen, E. De Coninck, B. Vankeirsbilck, P. Simoens, and B. Dhoedt, "Sensor fusion for robot control through deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 2365–2370.

[51] Google, "Tensorflow, a open source ml framwork." https://www.tensorflow.org/

[52] Google, "Keras, a high level python wrapper for tensorflow." https://keras.io/

[53] Facebook, "Pytorch, a open source ml framework." https://pytorch.org/

[54] Open Robotics, "About ros." https://www.ros.org/about-ros/

[55] E. Ackerman and E. Guizzo, "Wizards of ros: Willow garage and the making of the robot operating system." https://spectrum.ieee.org/automaton/robotics/robotics-software/wizards-of-ros-willow-garage-and-the-making-of-the-robot-operating-system

[56] B. Gerkey, "Why ros2?" http://design.ros2.org/articles/why_ros2.html

[57] Open Robotics, "Rosbridge suite." http://wiki.ros.org/rosbridge_suite

[58] Open Robotics, "Message filters." http://wiki.ros.org/message_filters

[59] Docker, Inc., "Docker - how to get started." https://docs.docker.com/get-started/

[60] LG SVL team, “Lgsvl simulator github page.” https://github.com/lgsvl/simulator

[61] Baidu, “Baidu-apollo.” https://github.com/ApolloAuto/apollo

[62] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, “An open approach to autonomous vehicles,” *IEEE Micro*, vol. 35, no. 6, pp. 60–68, 2015.

[63] Unity Technologies, “Unity.” https://unity.com/

[64] HDF Group, “The HDF5 Library & File Format.” https://www.hdfgroup.org/solutions/hdf5