# Problem Statement

The transportation sector is currently undergoing significant changes and developments, particularly in the past five years with the coming online of mobile apps and the development of technology platforms like Uber, Lyft and others. With the advent of mobile apps together with great advancement in machine learning including computer vision together with computing power and advancement have all encouraged the development, research and testing of autonomous vehicles.

Autonomous vehicles pose a variety of challenges – most importantly accurately identifying objects / cars on the roadway using inputs from sensor data. Data input from sensors can be processed to develop machine learning models, which can in turn be used to more accurately and rapidly deploy autonomous vehicles.

# Literature Review

Convolutional Neural Networks (CNN) have become an increasingly important tool in computer vision problems and have evolved rapidly in past few years in developing different structures to efficiently solve these problems. The section below is organized as follows, 2.1 provides a technical overview including the introduction of different concepts that are relevant for the approaches used, 2.2 provides a review of different network architectures that have been used in CNN’s, 2.3 provides a review of different approaches that have been adopted for conducting 3D object detection and processing Point Cloud (or Light Detection and Ranging – LiDAR) data and 2.4 provides an overview of the different datasets that have been used in the literature to develop models to solve this problem.

## 2.1 Technical Overview

The typical CNN’s have neurons arranged in 3 dimensions of width, height and depth as input, with the depth representing 3 color channels (RGB) for images. Compared to a typical neural network which consists of multiple layers of fully connected hidden layers, a CNN consists of a sequence of layers that convolve the input and yield an output i.e. convert the input to an output after performing a differentiable function on the input [1].

CNN’s have fewer or in some cases no fully connected layers and thus greatly reduce the number of parameters required to be estimated and learned during model training process. A simple CNN may consist of a convolution layer, a pooling layer and fully connected layer. The convolution layer typically has different filters which are applied to the input to extract different features of the output. For example, one filter may extract only vertical lines in the input, and another may extract only horizontal lines from the input. The number of filters associated with a convolutional layer and the size (width and height of the filter kernel) are hyperparameters for the convolutional network. The specific weights of the filter kernel are learned through training by minimizing a loss function similar to the case of fully connected layers [1]. Max pooling layers and are use to reduce the size of the input generally referred to as down-sampling, with idea being to identify key features of the input by reducing the size of the input.

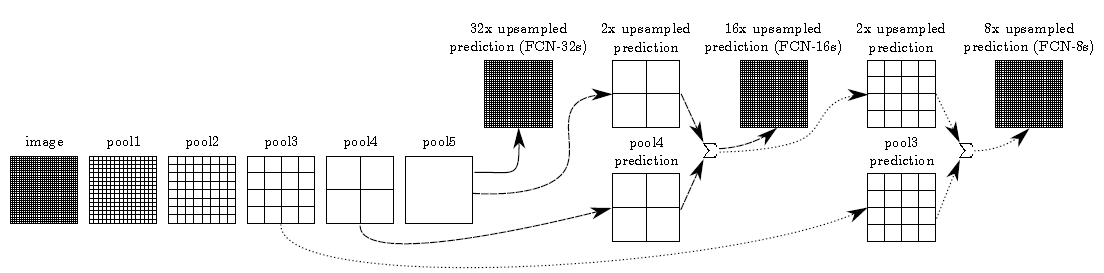
Convolutional layers typically reduce the size or down-sample the input. Deconvolution layers also referred to as transposed convolution layers take an input and give an output that that is of greater size. Deconvolutional layers can be used as decoding layer for a convolutional layer or to project feature maps to a higher dimensional space [2]. Deconvolution layers operate in a top-down fashion generating the input signal by combining convolutions of the feature maps with learned filters [3]. The approach allows for the unsupervised construction of hierarchical image representations. For example, in [6] the authors show that deconvolutional networks can learn different features which can be combined in the reconstructed image with object detection. Deconvolution layers are used extensively in the literature as part of the network architecture to extract higher level features from point cloud and image-based datasets.

Fully convolutional networks (FCN) are another adaption that employ convolutional networks for object detection and semantic segmentation [4,5]. FCN’s are unique in that they consist only of convolutional layers – convolution, pooling and up-sampling - and do not contain any fully connected hidden neural layer. This greatly reduces the number of parameters needed for computation of the network and thus are much faster than the typical convolutional network which has fully connected layers. In addition to this, a key insight of FCN’s is that any fully connected network can be converted to its corresponding FCN. The densely connected layer can be considered as a 1x1 convolutional layer [4]. In [4] the authors develop FCN versions of a number of convolutional networks such as LeNet, AlexNet and VGGNet.

FCN’s have two different path’s a down-sampling path (which reduces the size of the output during convolution) and an up-sampling path (increases the size of the output during convolution). The down-sampling path recovers local semantic and contextual information (the what) while the up-sampling path recovers the spatial information (the where). Up-sampling paths are deconvolutional layers while down-sampling paths represent typical convolutional layers.

To recover and combine the information learned through each path skip connections are used. Skip connections consists of connections that link the output of one layer to the input of another but skip or exclude in between layers so as to incorporate outputs of different layers concurrently. The figure below shows an example. FCN 32 is the output of a 5th layer of pooling (and convolution – convolution layers are not shown for brevity). FCN 16 is obtained by concatenating output from the 5th layer of pooling and a (skipped) 4th layer of pooling.

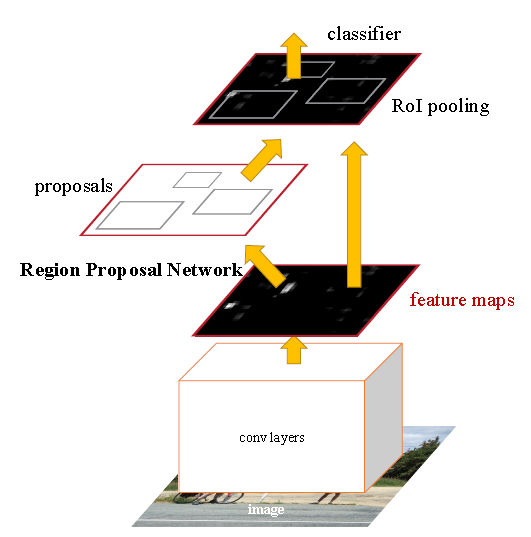
Figure : Skip connections for Semantic Segmentation

****

Source: [4]

RPN (Regional Proposal Networks) is a network that takes an image input of any size and simultaneously outputs rectangular object proposals together with a particular score for the object [17]. In [17] the authors introduce a novel RPN that shares convolution layers with an object detection network and thereby reducing time and cost for computing proposals. In simple words the RPN network tells the object detection network *where* to look in a particular image. The authors thus propose a two-stage approach, with a first stage of proposal generation followed by object identification whilst other approaches that conduct one stage identification of region and object detection. Based on the analysis, the two-stage approach has superior performance relative to the one stage approach. By using novel anchor boxes that use references at multiple scales and aspect ratios, the authors in [17], improve performance of the network (as shown in the figure below). The RPN is combined with the object detection network by a training scheme that alternates between loss reduction for the region proposal task and loss reduction for the object detection task, resulting in a unified network with convolutional features that are shared by the both tasks.

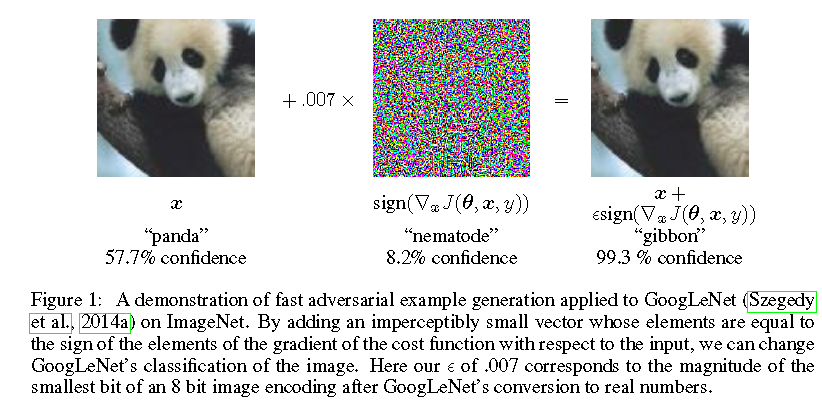
Figure : An RPN combined with a R-CNN



Source: [17]

In [7] the authors identify adversarial examples, where a small perturbation in the dataset examples results in neural networks misclassifying the output with a high degree of confidence. For example, in some cases such as ImageNet, the adversarial examples were so close to the original examples that the differences are indistinguishable to the human eye. The authors recommend training with adversarial examples which can result in regularization of the model.

Figure : An Adversarial Example applied to GoogLeNet on ImageNet



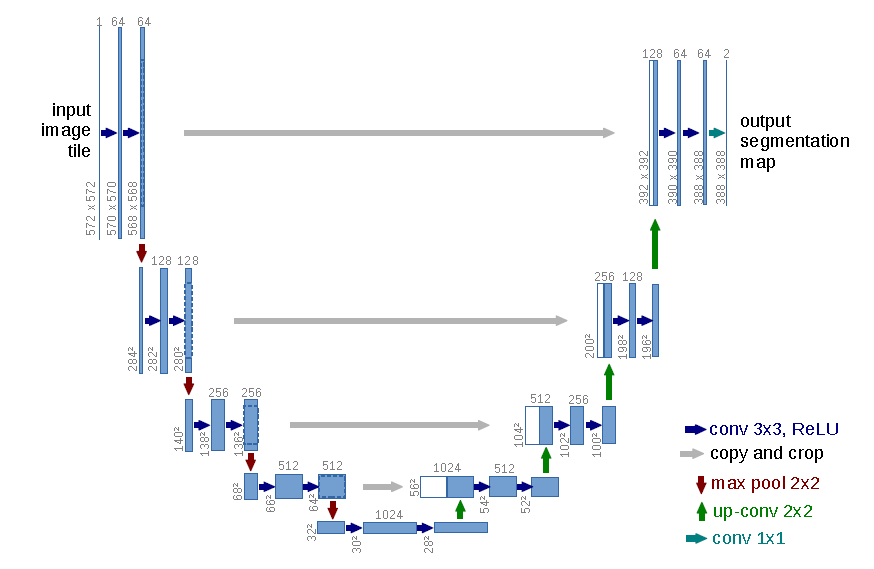
Source: [7]

## Network Architecture

The current state of the art references and uses a number of architectures. One of the main training architectures that are being used is the Unet [9]. The approach builds upon the FCN architecture that was discussed above. In [9], the authors adopt the FCN approach with some modifications that works with very few training images and provides precise segmentations. This approach won the 2015 International Symposium on Biomedical Imaging (ISBI). The figure below (taken from the paper) illustrates the approach and how it works. As can be seen in the figure the architecture consists of a contracting / down-sampling path and an expansive / up-sampling path. The down-sampling path consists of two convolutions followed by ReLU layers and a max-pooling layer. After each max pooling layer during the down-sampling phase the number of feature maps are doubled just as the size of output is down-sampled by a factor of two.

During the up-sampling phase, at each the level there are two convolutional layers followed by an up-convolution (synonymous with the deconvolutional layers we discussed above) the number of features are reduced by a factor of two just as the size of the output is increased (see Figure 4 below). A process of concatenation takes places that combines the up-sampled feature from the corresponding down-sampled features. The benefit of this concatenation is that during the down-sampling process very localized features are captured while during the up-sampling process higher level features are captured. At the final layer a 1x1 convolution is used to map the 64 component features vector to the desired number of classes.

Figure : The Unet Architecture



Source: [9]

Sparse CNN are another approach used to improve performance of convolutional networks and indeed to train deeper networks. In [10] the authors develop an approach to more efficiently represent data in sparse form. For forward propagation the authors calculate two matrices for each layer of the network, which consists of a feature matrix and a pointer matrix.

The feature matrix is a list of row vectors one for the base state (background for example) and one for the active spatial location in the layer. The number of features per spatial location is represented by width of the matrix. The pointer matrix is of size equal to the convolution layer. For each spatial location in the convolution layer, the number of corresponding in the feature matrix is stored.

The benefit of using sparse CNN’s is that they allow the training of faster networks and allow for the training of much deeper networks with comparable performance.

## Applications to 3D Object Detection and Point Cloud

This section presents approaches that have been covered in the sections above from a theoretical and application perspective but now specifically applied to applications for data that are applicable for autonomous vehicles. Two types of dataset are useful and applicable for autonomous vehicles that will be considered in this paper, namely image data and LIDAR (or point cloud data). There are other sources of data which will be briefly discussed but will not be considered for this analysis.

Image data provides good feature resolution and object detection capabilities however depth detection which is critical for 3D object detection and for applications in autonomous vehicles is not well represented. LIDAR data on the other hand provides good depth detection however it does have its own challenges. In particular, data from LIDAR is represented as a cloud of points (hence the name point cloud) with each frame having 100k or more points. This means the point locations need to be processed and can take up computational resources. In addition, approach speed of the vehicle that is hosting the LIDAR equipment can also impact the number of points per object available for processing. Finally, LIDAR points are sparse, have highly variable point density due to factors such as non-uniform sampling of the 3D space, effective range of sensors along with other factors and therefore require additional considerations.

Combining both LIDAR and image data together for processing may present a solution however the approach comes with additional challenges. For one, image capture cameras and LIDAR sweeps can become unsynchronized and so some additional processing is needed to ensure both sources are synchronized. Secondly, with such an approach, much more data needs to be processed which can slowdown the network and image segmentation / classification.

In the literature therefore four main approaches are generally used to deal with 3D object detection based using point cloud and image-based data [16].

The first consists of front view (camera) or image-based methods. These are generally based on 2D image representations of image data or Bird’s eye view (BEV). In front view methods, image-based methods are used to generate 2D bounding boxes including in some approaches utilizing CNN’s for analysis. In some approaches for analyzing front view LIDAR data are also used [14]. Generally, these methods do not perform well for 3D detection [16].

A second approach consists of first converting point cloud data into a BEV representation. In this case point cloud data are used to obtain height maps which are then combined with point intensity and density maps to obtain features [12]. However, one problem with these approaches is that there is a loss of information when developing the BEV map. This reduces performance of these models particularly for 3D bounding box applications.

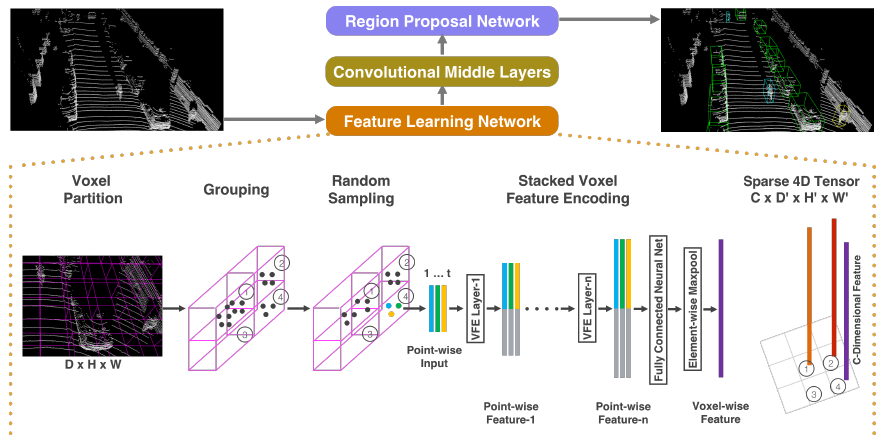
A third approach considers methods and applications in 3D. Methods in 3D generally work with LIDAR data (although there are other data sources such as stereo images) by converting point cloud data into 3D grids or voxels. In [13], the authors exploit the sparse nature of 3D data to develop a voting scheme that is equivalent to a convolution on a sparse feature grid. Using point cloud data they extract six features which include three measures based on the diffusion of the point cloud (these measures consider eigenvalues of the point cloud in 3D space, the mean and the variance of the reflectance values of the points and binary occupancy feature.

In [11] a generic 3D detection network is developed that conducts feature extraction as well as bounding box prediction into a single stage deep network. Instead of manual feature engineering this approach uses machine learned features. The authors divide the point cloud into equally spaced 3D voxels and transform a group of points within each voxel into a unified feature representation through a voxel feature encoding layer (VFE). The VFE layer then feeds into a RPN to generate detections. RPN’s are highly optimized for object detection however it requires data to be dense and organized in a tensor structure – which not typically the case for LIDAR which tends to be sparse.

As shown in the figure below the approach consists of 3 different steps: 1) feature learning network 2) convolutional middle layers and 3) regional proposal network (see Figure 5 below).

The feature learning network itself consists of a number of steps, these include as a first step creation of the voxel partitioning (essentially creating partitions of 3D space and grouping points according the partitioned space). Secondly, random sampling of LIDAR points from the partitioned voxel space. LIDAR is high definition consisting of ~100k points with highly variable point density. Random sampling provides computational savings and reduces the imbalance of points due to the variable point density.

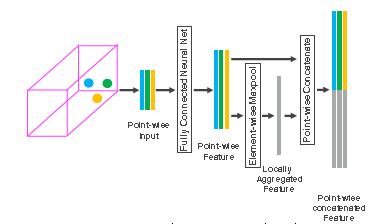
Figure : VoxelNET Approach

****

Source: [11]

Thirdly, stacked voxel feature encoding implements skip connections and concatenation to learning pointwise and locally aggregated features (see Figure 6 below). Fourthly, while LIDAR points consist of 100K points per LIDAR scene, about 90% of the voxels are empty, and the sparse tensor only considers non-empty voxels. Thus the sparse tensor representation greatly reduces memory usage and computation cost.

Figure : Stacked VFE Layers

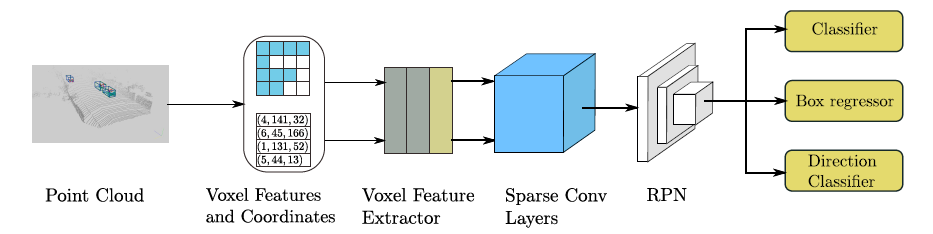


Source: [11]

The RPN in [11] consists of a number of convolutional layers that successively down-sampled the layers. Successive deconvolutional layers together with skip connections and deconvolutional layers up-sample the output which is concatenated and outputs a location and probability score map.

In [16] the authors only use LIDAR data to convert point cloud data into voxels and then applying linear networks eventually converting these into 3D tensors which are ultimately used in a region proposal network. The authors use a similar structure as [11] above with some notable changes including sparse convolution methods which greatly improve speed and training of these networks. In addition, in [16] the authors use an angle loss regression framework to improve orientation performance and data augmentation approaches to enhance convergence speed and performance.

Figure : SECOND Detector



Source: [16]

In [15] the LIDAR data is used on the Nuscenes dataset, using sparse 3D convolution to extract features which are then into a multi-task learning framework. The authors develop and apply a method to conduct multi-task learning by estimating and training all categories together rather training them individually.

Finally, a fourth approach considers a fusion based approach that combines both camera images with point cloud data. In [12] point cloud data are used to develop both front view feature maps and BEV based feature maps. These point cloud maps are then combined with an image feature map. The network performs better than the BEV only network however, performance is slower due to the model containing three CNN’s.

## Object Detection and Point Cloud Datasets

A number of different datasets are available the provide image and point cloud data for model training, testing and validation. The most well documented and extensive studied and used is the KITTI dataset [18]. The KITTI 3D object detection database consists of 7,481 training images / points clouds and 7,518 test images / point clouds covering three categories: car, pedestrian and cyclist. For each class detection, three difficultly levels are available for evaluation – easy, medium and hard. The categorization is based on object size, occlusion and truncation level [11]. For KITTI 3D object detection is based on LIDAR. On other hand Cityscapes [20] has released a dataset together with 3D bounding boxes with only RGB / stereo images (and not using LIDAR for depth perception).

Recently Nutonomy (the nuScenes dataset) released autonomous vehicle model training dataset and carries a full autonomous vehicle sensor suite, consisting of 6 cameras, 5 radars and 1 lidar providing 360 degree field of view [21]. NuScenes comprises of 1000 scenes (each 20 seconds long), fully annotated 3D bounding boxes for 23 classes and 8 attributes. In [21] devkit, evaluation code, taxonomoy and database schema are provided. NuScenes is a much more challenging dataset than KITTI, for one it requires detection of 10 categories simultaneously compared to 3 in the KITTI dataset. Additionally object classes in the dataset are significantly imbalanced making the nuScenes more challenging to work with.

Finally Lyft [19] has released a dataset together with a devkit and fully annotated 3D bounding boxes. The dataset consists of approximately 80 GB of data comprising of images, video and LIDAR.

For this study I propose to use the KITTI benchmark for model training, application and validation[[1]](#footnote-1).

# References

1. Li F, Krishna R., Xu D., Convolutional Neural Networks for Visual Recognition, Stanford 2020 (<https://cs231n.github.io/convolutional-networks/>)
2. Dumoulin V., Visin F., A Guide to Convolutional Arithmetic for Deep Learning, arXiv 2016 arXiv:1603.072285v1
3. Zeiler M., Krishnan D., Taylor G. and Fergus R. Deconvolutional Networks, CVPR 2010
4. Long J., Shelhamer E., Darrell T., Fully Convolutional Networks for Semantic Segmentation, arXiv 2015, arXiv: 1411.4038v2
5. Theano Development Team, 2008–2013, Fully Convolutional Networks for 2D Segmentation (<http://deeplearning.net/tutorial/fcn_2D_segm.html>)
6. Zeiler M., Taylor G., and Fergus R. Adaptive Deconvolutional Networks for Mid and High Level Feature Learning, International Conference on Computer Vision (ICCV) 2011
7. Goodfellow I., Shlens J., Explaining and Harnessing Adversarial Examples, International Conference on Learning Representation (ICLR) 2015
8. 2008–2013, Theano Development Team, Convolutional Neural Networks (LeNet) (<http://deeplearning.net/tutorial/lenet.html>)
9. Ronneburger O., Fisher P., Brox T., U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv 2015, arXiv 1505.04597v1
10. Graham B., Spatially Sparse Convolutional Neural Networks, arXiv 2014, arXiv: 1409.6070v1
11. Zhou Y. and Tuzel O., VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection, arXiv 2017, arXiv 1711.06396v1
12. Chen X., Ma H., Wan J., Li B., Xia T., Multi-view 3D Object Detection Network for Autonomous Driving, arXiv 2017, arXiv 1611.07759v3
13. Wang D., Posner I., Voting for Voting in Online Point Cloud Object Detection Proceedings of the Robotics: Science and Systems, Rome, Italy, 1317, 2015
14. Li B., Zhang T., Xia T., Vehicle Detection from 3D Lidar Using Fully Convolutional Network, arXiv 2016, arXiv:1608.07916v1
15. Zhu B., Jiang Z., Zhou X., Li Z., Gang Y., Class Balanced Grouping and Sampling for Point Cloud 3D Object Detection, arXiv 2019, arXiv: 1908.09492v1
16. Yan Y., Mao Y., Li B., SECOND: Sparsely Embedded Convolutional Detection, Sensors 2018
17. Ren S, He K., Girshcik R., and Sun J., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks arXiv 2016, arXiv: 1506.01497v3
18. Geiger A., Lenz P., Urtasun R., Are we ready for Autonomous Driving? The Kitti Vision Benchmark Suite, Computer Vision and Pattern Recognition (CVPR) 2012
19. Kesten R. ,Usman M.,Houston J., Pandya, T., Nadhamuni K., Ferreira A., Yuan M., Low B., Jain A., Ondruska P., Omari S. , Shah S., Kulkarni A.,Kazakova A.,Tao C., Platinsky L., Jiang W. and Shet, V., Lyft 2019, Lyft Level 5 Perception Dataset 2020, <https://level5.lyft.com/dataset/>
20. Cordts M., Omran M., Ramos S., Rehfeld T., Enzweiler M., Benenson R., Franke U., Roth S. and Schiele B., The Cityscapes Dataset for Semantic Urban Scene Understanding 2016.
21. Caesar H., Bankiti V., Lang A., Vora S., Liong V., Xu Q., Krishnan A., Pan Y., Giancarlo B., Beijbom O., nuScenes: A multimodal dataset for autonomous driving, arXiv 2020, arXiv 1903.11027v5

1. However, given the more detailed devkits that are available for Lyft and nuScenes I will explore these datasets as well. [↑](#footnote-ref-1)