A Survey on Semantic Segmentation using Deep Learning Techniques

B. Tapasvi[1]
PhD Scholar
ECE Department
Annamalai University
Chidambaram

Dr. N. Udaya Kumar[2]
Professor
ECE Department
SRKR Engineering College
Bhimavaram

Dr. E. Gnanamanoharan[3]
Assistant Professor
ECE Department
Annamalai University
Chidambaram

***Abstract*- Semantic segmentation is a challenging task in the field of computer vision. It is process of classifying each pixel belonging to a particular label. It has many challenging applications such as autonomous vehicles, human-computer interaction, robot navigation, medical research and so on, which motivates us to survey the different semantic segmentation architectures. Most of these methods have been built using the deep learning techniques. In this paper we made a review of some state-of-the-art Convolutional Neural Network(CNN) architectures such as AlexNet, GoogleNet, VGGNet, ResNet which form the basis for Semantic Segmentation. Further, we presenteddifferent semanticsegmentation architectures such as Fully Convolutional Network (FCN), ParseNet, Deconvolution Network, U-Net, Feature Pyramid Network(FPN), Mask R-CNN. Finally, we compared the performances of all these architectures.**

Keywords: **Semantic Segmentation, Deep Learning, Convolutional Neural Networks.**

1.INTRODUCTION

Semantic segmentation is one of the key problems in the field of computer vision. It plays an important role in image understanding and essential for image analysis tasks. It is a natural step in the progression from coarse to fine inference (bounding box to pixel wise representation). It refers to the process of linking each pixel in an image to a class label i.e., we can think of semantic segmentation as image classification at a pixel level [1]. The importance of scene understanding as a core computer vision problem is highlighted by the fact that an increasing number of applications nourish from inferring knowledge from imagery. It has several applications in computer vision & artificial intelligence such as autonomous driving, robot navigation, remote sensing, in agriculture sciences, in medical sciences (medical imaging analysis) etc. with the popularity of deep learning in recent years, many semantic segmentation problems are being tackled using deep architectures, most often by Convolution Neural Networks, which surpass other approaches by a large margin in terms of accuracy and efficiency.

Semantic segmentation is the process of inferring the progression from coarse to fine details in a scene. The process begins by performing classification which makes a prediction of the input. The next step is to do localization/detection to determine the spatial location of the detected classes. The process is completed when fine-grained inference is done based on dense predictions. The end result of semantic segmentation to label each pixel with class of its object in the region of interest.

The Figure 1 illustrates the semantic segmentation for pixel level classification, Figure 1(a) represents the ground truth image, the corresponding semantic segmentation image is represented in Figure 1(b) as it classifies the person, car, road, building, etc from the ground truth image similarly Figure 1(c) represents the ground truth image, the corresponding semantic segmentation image is represented in Figure 1(d) as it classifies the cow, grass, building, tree, etc from the ground truth image.

(a)

(b)

(c)

(d)

However, recent advances in computer vision & deep learning for semantic segmentation have made most of the older methods obsolete. Therefore, have turned towards deep learning architectures which are considered to be the state of the art and have achieved the top benchmark performance across the well-known international datasets. The latest semantic segmentation methods can be classified into three categories: Region based Semantic Segmentation, FCN based Semantic Segmentation and Weakly Supervised Segmentation.

In this paper a review on literature on the various techniques in semantic segmentation is presented. The layout of the paper is designed as follows: Section 2 describes the Deep Neural Network (DNN) architectures used in semantic segmentation. Section 3 describes the architectures of various semantic segmentation networks. Section 4 gives a comparison on performance of various semantic segmentation architectures and finally Section 5 concludes the paper.

2. DEEP LEARNING ARCHITECTURES FOR SEMANTIC SEGMENTATION

Generally, DNNs are the basis of semantic segmentation networks and have significant contribution towards the research in computer vision. So, the advances and developments in DNNs are discussed prior to the discussion of semantic segmentation architectures. The most popular and widely DNN architecture for computer vision related research are AlexNet, GoogleNet, VGGNet, ResNet etc.

a. AlexNet [2]

AlexNet is a network introduced by Alex Krizhevsky and is the winner of the ILSVRC 2012, which is an image classification competition, the dataset used is Image Net has over 15 million labelled high -resolution images with around 22000 categories. The architecture of Alex net contains eight layers in which five are convolution layers and three are fully connected layers as shown in Figure 2 and also consists of three pooling layers.

Before AlexNet, hyperbolic tangent

$$(\tanh=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}) \quad -(1)$$

was used as activation function. Rectified Linear Unit (ReLU) is introduced in AlexNet as activation function

$$f(x)=\begin{cases}0, & x<0\\ x, & x\geq 0\end{cases} \quad -(2).$$

ReLU is six times faster than tanh to reach 25% training error rate. Furthermore, data augmentation and dropout are widely used today as efficient learning strategies were first introduced in Alexnet. Hence it is known as the foundation work of modern deep convolution neural network (CNN).

b. GoogleNet [3]

GoogleNet was proposed by Christian Szegedy andis the winner of the ILSVRC 2014 competition was GoogLeNet(a.k.a. Inception V1) from Google. The network used a CNN inspired by LeNet but implemented a novel building block which is named as inception module. This module is based on several very small convolutions in order to drastically reduce the number of parameters.This architecture consists of a 22-layer deep CNN but reduced the number of parameters from 60 million (AlexNet) to 4 million.

c. VGGNet [4]

VGGNet is first introduced by Karen Simonyan and is the runner-up at the ILSVRC 2014 competition. It consists of 16 convolutional layers and is very appealing because of its very uniform architecture. Similar to AlexNet, only 3x3 convolutions, but has more filters. VGGNet has gained its popularity in the research community for extracting features from images because it uses a stack of convolution layers with small receptive fields in the first layers instead of few layers with big receptive fields.The weight configuration of the VGGNet is publicly available and has been used inmany other applications and challenges as a baseline feature extractor.

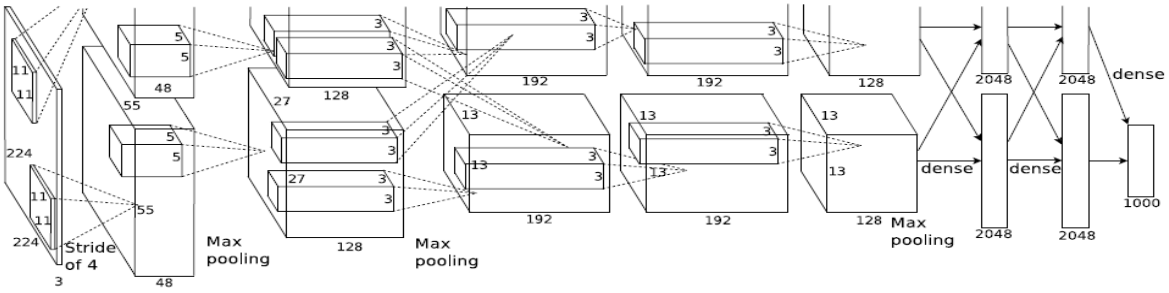

Figure 2: AlexNet Architecture

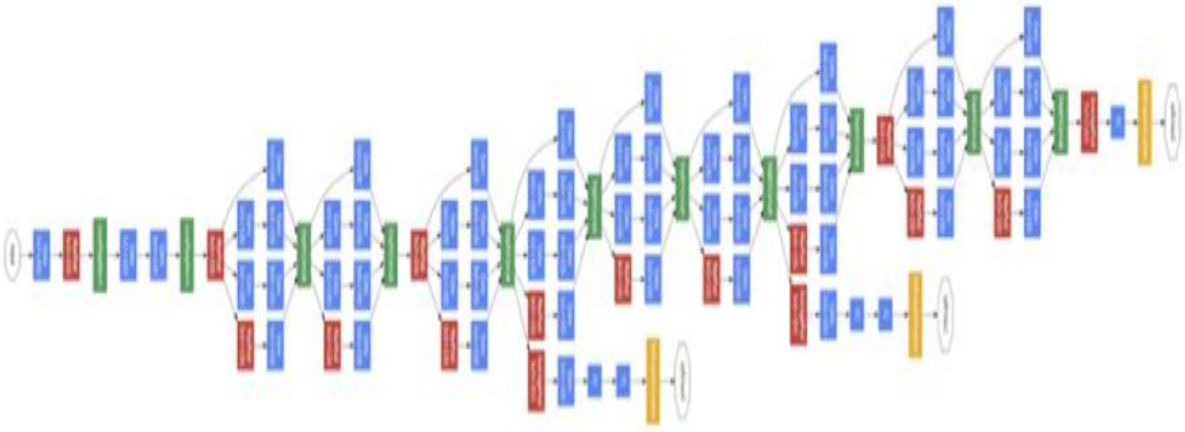

Figure 3: GoogleNet Architecture

Figure 4: VGGNet Architecture

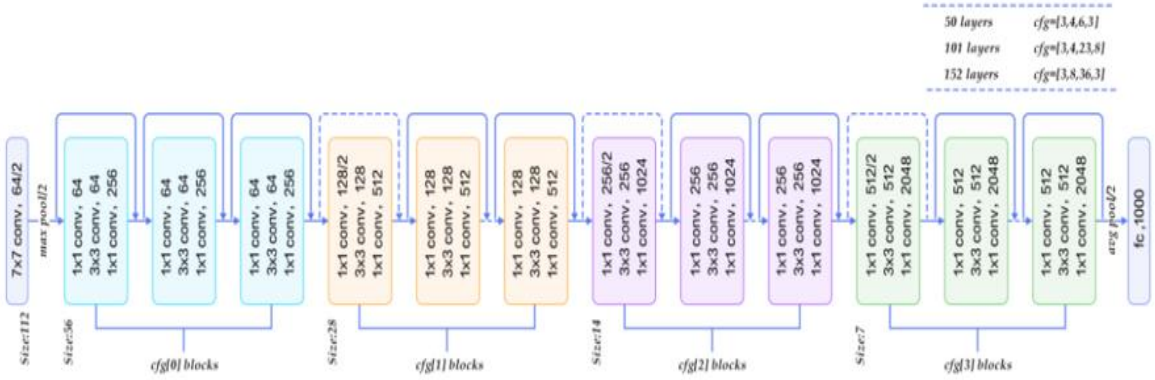

Figure 5: ResNet Architecture

However, VGGNet consists of 138 million parameters, which can be a bit challenging to handle.

d. ResNet [5]

ResNet was proposed by Kaiming He and is the winner at the ILSVRC 2016 with 96.4% accuracy. Kaiming He et al introduced a novel architecture with "skip connections" and features heavy batch normalization. Such skip connections are also known as gated units or gated recurrent units and have a strong similarity to recent successful elements applied in Recurrent Neural Networks (RNN). The intuitive idea behind this approach is that the next layer learns something new and different from what the input has already learned. This technique is able to train a Neural Network with152 layers while still having lower complexity than VGGNet. It achieves a top-5 error rate of 3.57% which beats human-level performance on this dataset.

Table 1: Summary of different CNN models on ImageNet classification task.

Model	Time	Accuracy	Number of parameters	Number of Layers
AlexNet	2012	57.2%	60M	8
VGGNet	2014	71.5%	138M	16
GoogleNet	2014	69.8%	6.8M	22
ResNet	2015	78.6%	55M	152

Till now DNN architectures which form the basis of semantic segmentation are discussed. In the next section the semantic segmentation architectures will be discussed.

3. Semantic Segmentation Architectures

The most popular and widely semantic segmentation architecture for computer vision related research are fully convolutional network(FCN), ParseNet, deconvolution network, U-Net, feature pyramid network(FPN), mask R-CNN etc.

a. Fully Convolutional Network for Semantic Segmentation [1]

Fully Convolutional Network(FCN) is developed by Jonathan Long, it is first of kind first to develop a Fully Convolutional Network(FCN) trained end-to-end for image segmentation. One of the best properties of FCN is, it takes an image with arbitrarysize and produces a segmented image with the same size. FCN is a network consisted of only convolutional layers as shown in figure 6. The authors start by modifying proven architectures such as AlexNet, VGGNet, GoogLeNet to have a nonfixed size input while replacing all the fully connected layers by convolutional layers and to retain the spatial information skip connections are introduced in the architecture. Since the network produces several feature maps with small sizes and dense representations, an upsampling is necessary to create an output with the same size as the input. It is commonly called deconvolution. They have also added skip connections in the network to combine high level feature map representations with specific and dense ones at the top of network. The model proposed in this paper achieves a performance of 67.2% mean Intersection over Union on PASCAL VOC 2012 dataset.

b. ParseNet [6]

ParseNet was proposed by Wei Liu in which explaining improvements of the FCN model. In this paper mentions that the FCN model loses the global context of the image in its deep layers. Context is known to be very useful for improving performance on detection and segmentation tasks using deep learning. From the Figure 7, the architecture having the convolution layer and normalization using l2 norm is performed for each channel at the lower path. At the upper path, convolution layer with global average pooling of those feature maps and perform l2 norm normalization. The ParseNet is an end-to-end convolution network predicting values for all the pixels at the same time and it avoids taking regions as input to keep the global information. It has obtained a 40.4% mean IoU score on the PASCAL Context challenge and a 69.55 mean IoU score on the PASCAL VOC segmentation challenge.

c. Learning Deconvolution Network for Semantic Segmentation [7]

Hyeonwoo Noh proposed a novel semantic segmentation algorithm by learning a deconvolution network. It is an end-to-end model composed of two linked parts. The first part is a convolutional network with a VGG16 architecture. The second part is a deconvolutional network which is novel part in this architecture, taking the vector of features as input and generating a map of pixel-wise probabilities corresponding to each class. Deconvolution is just to convolve input back to its original size. This network has obtained a 72.5% mean IoU on the 2012 PASCAL VOC segmentation challenge.

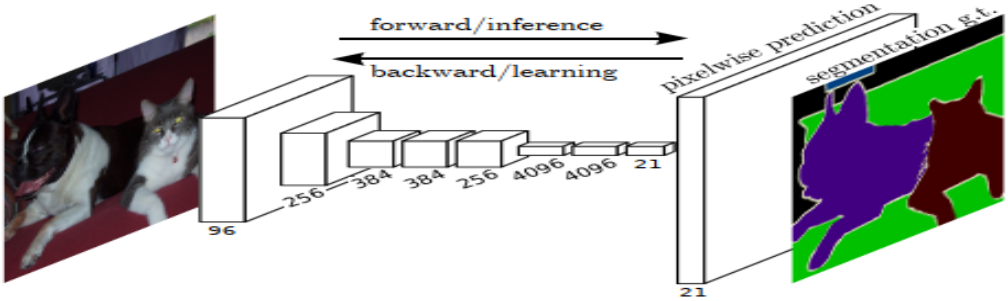

Figure 6: Fully Convolutional Network Architecture

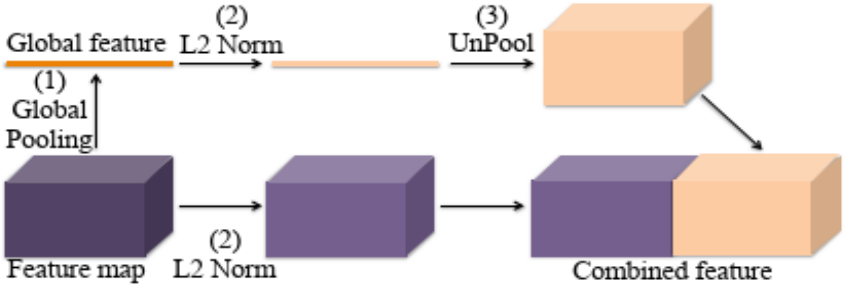

Figure 7: ParseNet Architecture

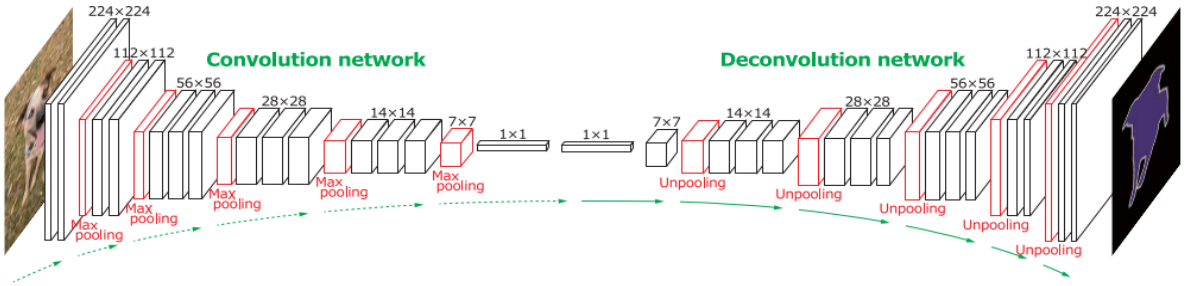

Figure 8: Deconvolution Network Architecture

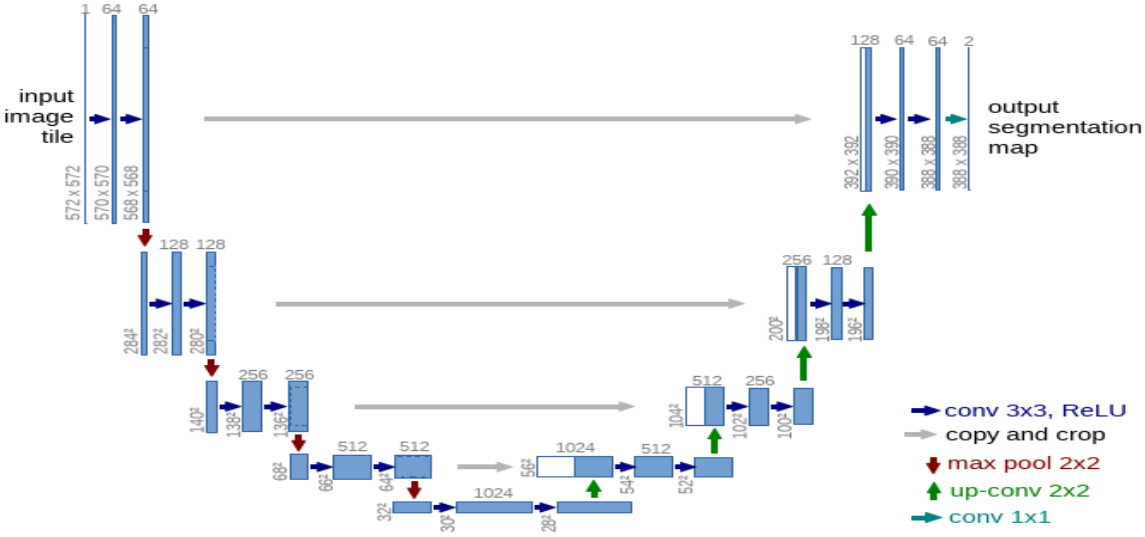

Figure 9: U-Net Architecture

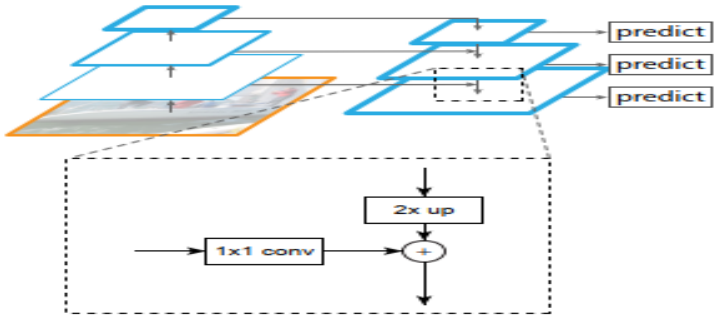

Figure 10: A building block of FPN

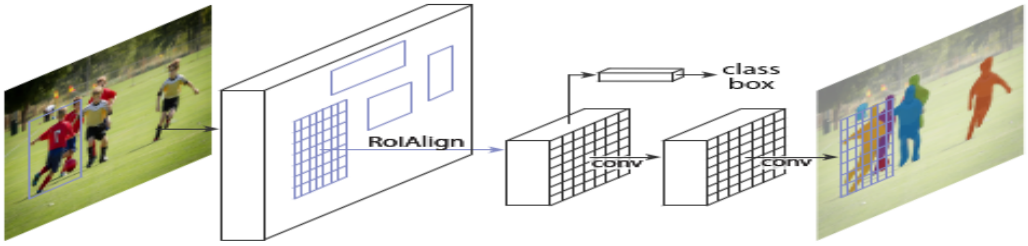

Figure 11: Mask R-CNN Architecture

d. U-Net [8]

OlagRonneberger proposed the U-Net architecture which extended the FCN architecture for biological microscopy images. It consists of two parts, a contracting part and an expanding part as shown in Figure 9. Contracting part is to compute features and expanding part is for spatially localise patterns in the image. The downsampling or contracting part has a FCN-like architecture used for extracting features with convolutions. The upsampling or expanding part uses deconvolution reducing the number of feature maps while increasing their height and width. The U-Net architecture consists of the repeatedapplication of two 3x3 convolutions (unpadded convolutions), each followed bya rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2for downsampling. At each downsampling step we double the number of featurechannels. Every step in the expansive path consists of an upsampling of thefeature map followed by a 2x2

convolution that halves thenumber of feature channels, a concatenation with the correspondingly croppedfeature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU.Compared to FCN-8, the two main differences are, U-net is symmetric and the skip connections between the down sampling path and the up sampling path apply a concatenation operator instead of a sum.A major advantage of U-net is that it is much faster to run than FCN or Mask RCNN.

e. Feature Pyramid Network (FPN) [9]

Tsung-Yi Lin have developed the Feature Pyramid Networks for Object Detection in 2016. Its architecture is composed of a bottom-up pathway, a top-down pathway and lateral connections in order to join low-resolution featuresas shown in figure10. The bottom-up pathway takes an image with an arbitrary size as input, processed with convolutional layers and downsampled by pooling layers. The top-down pathway consists of upsampling process. The FPN architecture achieves 48.1% Average Recall(AR) score on the 2016 COCO segmentation challenge.

f. Mask R-CNN [10]

Kaiming He have released the Mask R-CNN model beating all previous benchmarks on many COCO challenges. The Mask R-CNN is a Faster R-CNN with 3 output branches: the first computes the bounding box coordinates, the second one computes the associated class and the last one computes the binary mask to segment the object. From figure 11, the architecture consists of the convolutional backbone architecture used for feature extraction over an entire image, and the network head for bounding-box recognition (classification and regression) and mask prediction that is applied separately to each RoI. It has obtained a 37.1% AP score on the 2016 COCO segmentation challenge and a 41.8% AP score on the 2017 COCO segmentation challenge.

4. PERFORMANCE COMPARISON OF DIFFERENT SEMANTIC SEGMENTATION

All these semantic segmentation architectures discussed above are compared using some parameter metrics, here considered are mean intersection over union(mIoU), average precision (AR), average recall(AR).

Intersection over union(IoU):

The IoU is given by the ratio of the area of intersection and area of union of the predicted bounding box and ground truth bounding box.

Precision and Recall:

In the field of statistics and data science, precision of a given class in classification, a.k.a. positive predicted value, is given as the ratio of true positive (TP) and the total number of predicted positives i.e., (TP +FP (false positive)).

Similarly, the recall, a.k.a. true positive rate or sensitivity, of a given class in classification, is defined as the ratio of TP (true positive) and total of ground truth positives (TP+FN (false negative)).

Table 2: Overview of Scores of the Models Over the Different Datasets

Model	2012 PASCAL VOC (mIoU)	PASCAL-Context (mIoU)	2016 COCO (AP)	2016 COCO (AR)	2017 COCO (AP)
FCN	67.2	x	x	x	x
ParseNet	69.55	40.4	x	x	x
Conv&Deconv	72.5	x	x	x	x
U-Net	x	x	x	x	x
FPN	x	x	x	48.1	x
Mask R-CNN	x	x	37.1	x	41.8

Semantic segmentation techniques using deep neural networks are rapidly growing but having some problems.

It is required to reduce complexity and computation. DNNs require high memory consumption and time so these are not suitable for mobile devices which are having limited resources. Also, there is a problem with computation complexity that arises due to high number of operators which are due to involving of high number of parameters. So, it is important to investigate how to reducethecomplexity to achieve high performance without loss of accuracy

These semantic segmentation techniques need large and high-quality labelled data. Current state of the art methods requires high quality labelled data, which is difficult on large scale data set because it is more laborious and time consuming. The effective solution is to build large and high-quality datasets which is hard to achieve. So, to this problem the researchers rely on weakly supervised methods.

As DNNs require large data, they did not perform well unless they are given with large data sets. So, there may be a problem of overfitting. It occurs when the gap between the training error and test error is very large. Regularization techniques help in overcoming this problem. Some of the methods which are applied in DNNs to prevent overfitting such as L1 & L2 regularization, dropout, etc. Data augmentation is also used for reducing overfitting by increasing the size of the training data (image rotating, flipping, scaling, shifting).

Real time implementation of semantic segmentation techniques is also very important, as it can useful in autonomous cars, robot interaction etc., where the time consumption is also a crucial parameter while evaluating the performance of the system. One possible solution to the time consumption is performing the convolution operations in an efficient way in order to make the network having less parameters.

Therefore, for the better performance of the techniques it is necessary

1. To improve computational efficiency
2. To improve accuracy by eliminating background noise.

5. Conclusion

In this paper, we have provided a detail survey of deep learning techniques used for semantic segmentation and compared their performances by means of some parameter metrics such as mean IoU, AP and AR. This survey shows that there is much scope of improvement in terms of accuracy, speed and complexity. It is observed that it is required to reduce complexity and computation and there is a problem of overfitting. Real time implementation of semantic segmentation techniques is also very important. So, our future work will be to take some of these methods and develop a new one which can able to improve computational efficiency and also improve accuracy by eliminating background noise.

REFERENCES

[1] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[2] , Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *NIPS*. 2012.

[3] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[5] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[6] Liu, Wei, Andrew Rabinovich, and Alexander C. Berg. "Parsenet: Looking wider to see better." *arXiv preprint arXiv:1506.04579* (2015).

[7] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." *Proceedings of the IEEE international conference on computer vision*. 2015.

[8] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

[9] , Tsung-Yi, et al. "Feature pyramid networks for object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[10] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.