11-785 FINAL REPORT

SAD - SEMANTIC SEGMENTATION FOR AUTONOMOUS DRIVING

Gagan Shivananda Bharathi Gangothri
gshivana@andrew.cmu.edu

Janvi Jatakia
jjatakia@andrew.cmu.edu

Rahul Ramakrishnan
rramkris@andrew.cmu.edu

Shailee Vora
skvora@andrew.cmu.edu

ABSTRACT

The aim of our project is to use semantic segmentation in real life scenarios to support applications like autonomous driving. Thus, we have chosen a synthetic dataset of urban images named SYNTHIA for this purpose. We intend to use Deep Convolutional Neural Networks (DCNNs) for this visual-based semantic segmentation of urban images as DCNNs outperform reliable classifiers for such visual tasks as per recent findings. Our goal is to make a robust algorithm for semantic segmentation of urban images which can be trained on synthetic datasets but are scalable to real world scenes for testing.

1 INTRODUCTION

There has been a major increase in the development of autonomous vehicles recently. One major concern that comes up with autonomous cars is that the prediction of objects across the roads needs to be precise, because the lives of people are at risk here. Also, for any new startup that wants to enter the autonomous vehicle industry, having loads of data for training are very important but its very difficult to have thousand miles of data which makes it tough for them to compete with leading companies who have million miles of data. Having an algorithm that trains on synthetic data that can be generated at their will but works for real life scenes would be their lifeline to compete against world leaders. We have planned to work on making a robust model that will use semantic segmentation to classify the objects in the scene or to parse a scene. Semantic Segmentation in the case of images is pixel-wise detection of the images to detect edges and classify the objects. For making our model, we will be using the SYNTHIA datasetRos et al. (2016) that has images of the road scenarios. According to Ros et al. (2016), there have been various Deep Convolutional Neural Network models that can be implemented for the current problem. The fully connected neural network model Hoffman et al. (2016), had some very good results, hence we are planning to use DCNNs as our approach and baseline.

2 EXPERIMENTS

Our baseline assessment from the start of the project till mid-way was to achieve around 62% of accuracy using the FCN model as introduced in Long et al. (2014). However, further in our research

and literature survey we found a pyramid Scene Parsing (PSPNet) as mentioned in Zhao et al. (2016) as a better architecture. We learned PSPNet outperforms FCN models with a 82% accuracy on CityScapes dataset. This can be accredited to the following observations.

FCN versus PSPNet:
a. Accuracy of any scene parsing strategy could be based on how well the strategy captures the global context and utilize those clues to label objects and categorize each pixel. FCN models lack such strategy.

On the other hand, PSPNet captures global context by using large pooling layers and provide required attention to global clues and at the end, fuse different levels of features to make the predictions more reliable. We experimented this by testing the same image over FCN and PSP models in an interactive online visualizer which clearly showed us that FCN has a poor performance over PSPNet. FCN had a lot of mislabeled pixels unlike PSPNet.

b. PSPNet better accommodates objects of different sizes in the scene than FCNs.

c. Finally, PSPNet is better suited for complex scene understanding and dealing with variety of labels. This is directly evident by the fact that PSPNet outperforms FCNs and the state-of-the-art architecture for PASCAL VOC 2012 and Cityscapes dataset.

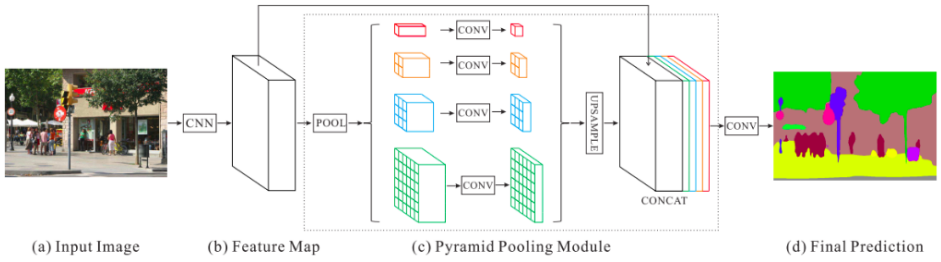

Fig 2.1 PSPNet Architecture

After carrying out further research, we found a new semantic segmentation model. This is developed by Google in Tensorflow known as DeepLabv3+. This method is a combination of ResNets, Atrous Convolutions and Atrous Spatial Pyramid Pooling. Atrous Convolution is also called as the dilated convolution. This model hasn't been used on the Synthia Dataset and also have many reasons of giving a better result hence, we also planned on studying it further. Following are some of the observations we made.

a. Dilated convolution means that the field of vision of the convolution filter is increased. This means that a global context is being considered which will help detection better.

b. The atrous spatial pyramid pooling works similar to the pooling in PSP net but instead of increading the filter size, just the field of view is increased, thus making the model train faster.

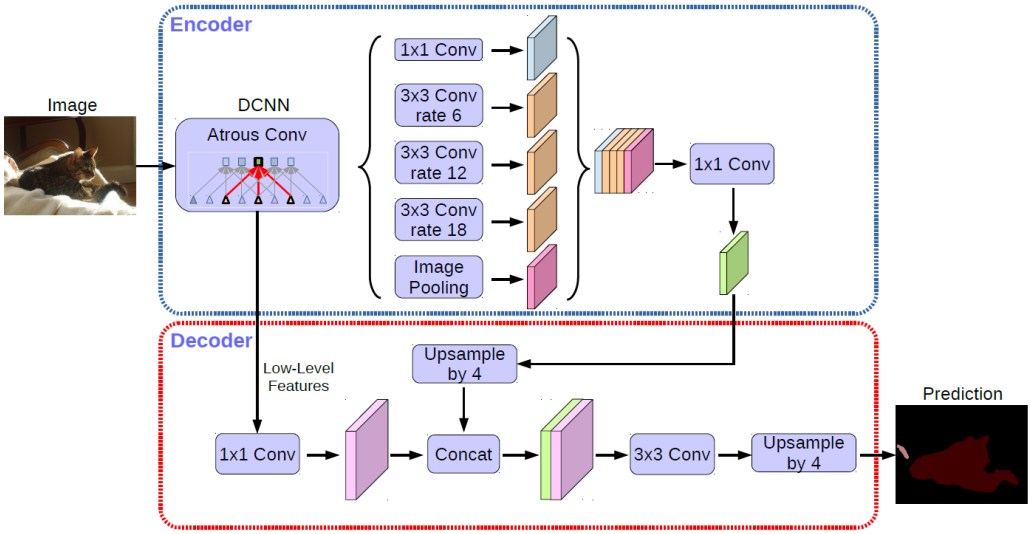

Fig 2.2 DeepLabv3+ Architecture

3 Proof of Concept

We trained two different models with completely different approach one being PSPNet and the other being DeepLabV3+. Both the models were trained on the synthia rand-cityscapes dataset containing 9400 images until they overfit. PSPNet started to overfit in 30 epochs whereas DeepLabV3+ started to overfit around 74 epochs.

After training, the models were tested on test images of synthia to validate if the training was successful. Once the results were satisfactory, we tested the same model without any modifications on real-world images from cityscapes to validate our model's accuracy. We compared the results of our best model among PSPNet and DeepLabV3+ with the state of the art paper AdaptationSeg published in ICCV 2017.

4 Code Implementation

4.1 Dataset Class

The SYNTHIA dataset has 23 classes. We first made a palette for the classes. This palette consists of the RGB values for each of the classes. Transformations are then applied to the images like cropping, etc and according to the input from the train file. The output of the dataset class would be images and the corresponding mask of the same size, with class assigned to each pixel.

4.2 PSPNet

A pyramid scene parsing network is basically proposed to embed difficult scenery context features in an FCN based pixel prediction framework. It is an effective optimization strategy developed for deep ResNet based on deeply supervised loss. We found a basic code from the original authors of the paper coded in Caffe. Since we were more comfortable coding in Pytorch and were confident enough to handle its intricacies, we used the Caffe model as a pseudo-code and created an Pytorch equivalent of it with added optimizations. Our architecture was ResNet-101. The custom dataset we created worked well and the training was successful. The model started to overfit around 30th epoch. We used the model saved at 30th epoch for performing our final testings.

4.3 DeepLabV3+

DeepLab has the structure of encoder and decoder. It was originally written in tensorflow and we decided to go with that rather than building our own code from scratch. We had to make significant

modifications to the source code to make it work for our dataset. The original code was written for resnet-50 and we urgraded that to use resnet-152. The traning was successful, the model started to overfit around 74th epoch. We used the model saved at 74th epoch for our final testing.

4.4 Important Considerations

- **Mismatched Relationship**: Context relationship is universal and important especially for complex scene understanding. There exist concurrent visual patterns. For example, an airplane is likely to be in runway or fly in sky while not over a road.

- **Confusion Categories**: Similar or confusion categories should be excluded so that the whole object is covered by a single label and not multiple labels. This problem can be remedied by utilizing the relationship between categories.

- **Inconspicuous Classes**: To improve performance for remarkably small or large objects, one should pay much attention to different sub-regions that contain inconspicuous classes.

5 Results and Validation

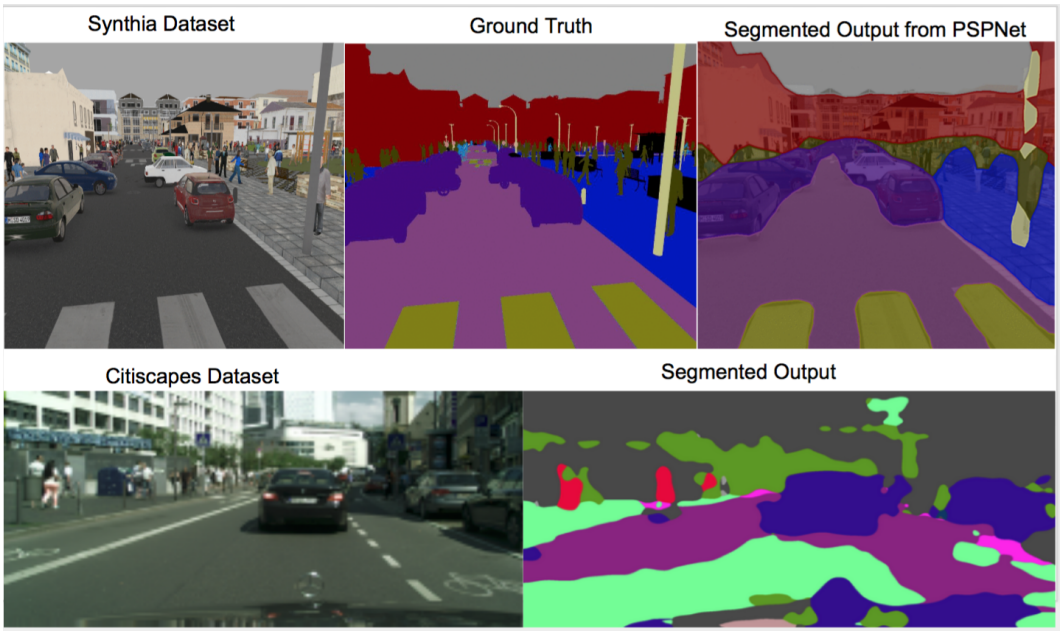

Fig 5.1 PSPNet Results

As visible from Fig 5.1, the ground truth and the output of the PSPNet model look reasonably similar. However, the lines between any two classes or segments of image are a little fuzzy. Similarly, on testing the model on real world data, we get a good result but not up to the mark.

In Fig 5.2, the ground truth and the output of the DeepLabv3+ look extremely similar. The lines between different classes also clearly distinguish the classes. This is a major improvement over the PSPNET model. Moreover, on testing the model on real world data, we get very satisfying results.

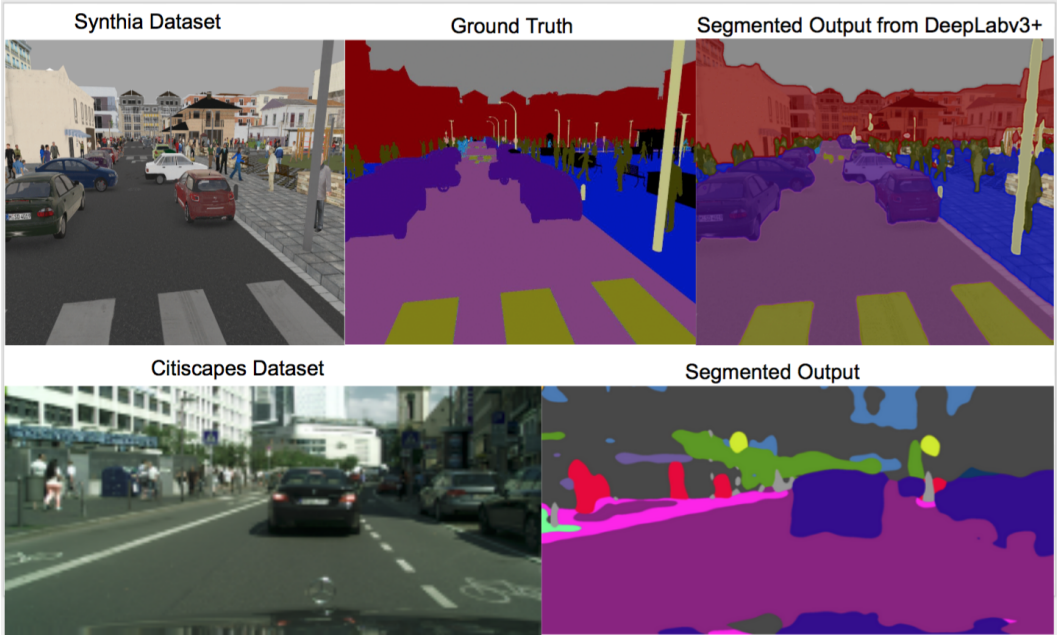

Fig 5.2 DeepLabv3+ Results

The reason for a better result here is that while training PSPNet the model started overfitting after 30 epochs whereas while training DeepLabv3+ the model started overfitting only after 74 epochs.

Given below is the evaluation of both our models on real world data called Citiscapes. We have used mean Intersection Over Union (IoU) as our evaluation metric. Intersection Over Union is calculated as the ratio of the intersection of the pixels in the predicted output and the ground truth (i.e. the correctly predicted pixels) and the union of all the pixels in the predicted output and the ground truth.

IoU on class-level	
Mean IoU	35.2
Road	90.8
Sidewalk	54.3
Building	71.6
Wall	14.5
Fence	13.4
Pole	22
Trafficlight	10.9
Trafficsign	15.7
Vegetation	78.1
Terrain	35.5
Sky	83.6
Person	33.5
Rider	9
Car	72.6
Truck	10.8
Bus	12.3
Train	0.8
Motorcycle	4.6
Bicycle	35.1

IoU on category-level	
Mean IoU	65.7
Flat	95
Nature	78
Object	22.9
Sky	83.6
Construction	71.6
Human	37.4
Vehicle	71.1

Fig 5.3 Mean IoU values at class-level and category-level on using the PSPNet model

IoU on class-level	
Mean IoU	49.3
Road	94.6
Sidewalk	65
Building	81.2
Wall	28
Fence	28.8
Pole	31.4
Trafficlight	26.5
Trafficsign	36.8
Vegetation	82.3
Terrain	46
Sky	86.5
Person	51.6
Rider	30.9
Car	82.9
Truck	32.8
Bus	38.2
Train	27.5
Motorcycle	15.3
Bicycle	49.6

IoU on category-level	
Mean IoU	74.1
Flat	96.5
Nature	82.4
Object	36.2
Sky	86.5
Construction	81.2
Human	54.6
Vehicle	81.3

Fig 5.4 Mean IoU values at class-level and category-level on using the DeepLabv3+ model

As you can see from Fig 5.3 and Fig 5.4 , the mean IoU of DeepLabv3+ is much better than that of PSPNET. Thus, we now have numerical evidence to conclude that the DeepLabv3+ model is better than PSPNet model.

Next, we compared our results with the state-of-the-art research paper in this domain called AdaptationSeg (ICCV 2017). The figure below gives the comparison between the mean IoU obtained by the author's of that paper and the mean IoU we obtained using our better model, DeepLabv3+. The first column has results of AdaptationSeg and the second column has results from DeepLabv3+.

IoU on class-level		
Mean iIoU	57	49.3
Road	96.4	94.6
Sidewalk	73.2	65
Building	84	81.2
Wall	28.5	28
Fence	29	28.8
Pole	35.7	31.4
Trafficlight	39.8	26.5
Trafficsign	45.2	36.8
Vegetation	87	82.3
Terrain	63.8	46
Sky	91.8	86.5
Person	62.8	51.6
Rider	42.8	30.9
Car	89.3	82.9
Truck	38.1	32.8
Bus	43.1	38.2
Train	44.2	27.5
Motorcycle	35.8	15.3
Bicycle	51.9	49.6

IoU on category-level		
Mean IoU	79.1	74.1
Flat	97.4	96.5
Nature	86.7	82.4
Object	42.5	36.2
Sky	91.8	86.5
Construction	83.8	81.2
Human	64.7	54.6
Vehicle	87.2	81.3

Fig 5.6 Mean IoU values at class-level and category-level using AdaptationSeg versus using our DeepLabv3+ model

From Fig 5.6 we can see that AdaptationSeg has better results than what we got. However, the reason for this could be attributed to the fact that the authors of AdaptationSeg, first trained on the synthetic dataset, froze the weights, and then trained on 500 real world samples. This gave their results an

edge over ours. We believe that if we also train on real world data then we can get comparable results.

6 Related Work

Semantic Segmentation is one of the most popular tasks in computer vision which is gaining a lot of traction especially after advent of the driver-less cars concept.

In the paper Long et al. (2014), Jonathan Long et al, show that a fully convolutional network (FCN) trained end-to-end, pixels-to-pixels on semantic segmentation exceeds the state-of-the-art without further machinery back then. and their goal is to make the process efficient by enabling pixel-wise prediction during upsampling layers and learning in nets with subsampled pooling.

Mottaghi et al. (2014) emphasize the importance of the role of the context in object detection and semantic segmentation. to analyse the effect of contexual information in detecting and segmenting images. The intuition behind this is that humans are worse than machines at classifying small image patches but are far better when more contextual information is available.They label each pixel of the training set of PASCAL to analyse the same.

Arbelaez et al. (2012) focusing on segmentation of objects particularly humans and animals introduce a pixel level classification approach as follows. They train the final set of classifiers that operate on pixels rather than on regions. This is how they generate feature vectors per pixel considering the following ways of projecting region scores onto pixels: Each pixel receives the average score of all the regions it is part of.Each pixel receives the maximum score among all the regions it is part of.They do non-max suppression on the regions, choosing the highest scoring region, then discard all overlapping regions, and repeat. Each pixel then receives the score of the highest scoring region.

Yao et al. (2012) This paper adresses the problem of scene understanding. It thus classifies the scene type i.e. city, sea, etc as well as the presence of various objects in those scenes which means car, boat, building dog, etc. They use a message passing scheme to transfer information to make the inference possible. The inference is done using the conditional random fields CRFs that contain a segment and a supersegment layer. The segments and super-segments show class labels to be assigned to each pixel in the image.

Finally, In the paper Hoffman et al. (2016), Hoffman et al, focus on the problem of domain shifts that are apparent in the fully convolutional models. For example, training on one city and testing on another in a different geographic region and/or weather condition may result in significantly degraded performance due to pixel-level distribution shift They propose an unsupervised adversarial approach to pixel prediction problems and work on both global and category specific adaptation techniques. They base their evaluation methods on SYNTHIA and GTA5 large datasets. They study a medium sized shift, through adaptation across season patterns observed within the SYNTHIA datasetHoffman et al. (2016). They also explore situations of relatively smaller domain shift within the CityScapes dataset.We are planning base our evaluation along the similar lines.

7 Data & Technical Requirements

We are using the SYNTHIA-RAND-CITYSCAPES dataset for training and validation. This dataset has automatically generated class annotations. We will test on some unseen data from the SYNTHIA dataset. Eventually, we intend to scale our system to perform semantic segmentation of real world urban images. The manually provided annotations of these real world images are what we will compare our generated labels against to check how well our model is performing.

References

Pablo Arbelaez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir D. Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *CVPR*, pp. 3378–3385. IEEE Computer Society, 2012. ISBN 978-1-4673-1226-4. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2012.html#ArbelaezHGGBM12.

Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. URL http://arxiv.org/abs/1612.02649.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. URL http://arxiv.org/abs/1411.4038.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. 2016.

Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pp. 702–709. IEEE Computer Society, 2012. ISBN 978-1-4673-1226-4. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2012.html#YaoFU12.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *CoRR*, abs/1612.01105, 2016. URL http://arxiv.org/abs/1612.01105.