Semantic Segmentation With Multi Scale Spatial Attention For Self Driving Cars

Abhinav Sagar, RajKumar Soundrapandiyan

Vellore Institute of Technology

Abstract

In this paper, we present a novel neural network using multi scale feature fusion at various scales for accurate and efficient semantic image segmentation. We used ResNet based feature extractor, dilated convolutional layers in downsampling part, atrous convolutional layers in the upsampling part and used concat operation to merge them. A new attention module is proposed to encode more contextual information and enhance the receptive field of the network. We present an in depth theoretical analysis of our network with training and optimization details. Our network was trained and tested on the Camvid dataset and Cityscapes dataset using mean accuracy per class and Intersection Over Union (IOU) as the evaluation metrics. Our model outperforms previous state of the art methods on semantic segmentation achieving mean IOU value of 74.12 while running at >100 FPS.

Introduction

- We propose a new model architecture which used dilated convolutional layers in downsampling part and atrous convolutional layers in upsampling at multiple scales.
- Concat operator is used for merging the feature maps for context encoding. We also propose our very own attention module which encodes channel wise information to model more contextual information and enlarge the receptive field. • We present the layer wise details, optimization and ablation study of our neural network.
- On evaluating our network using Camvid dataset and Cityscapes dataset using mean accuracy per class and IOU as evaluation metrics, our model outperforms previous state of the art model architectures while running at > 100 FPS.

Attention Module

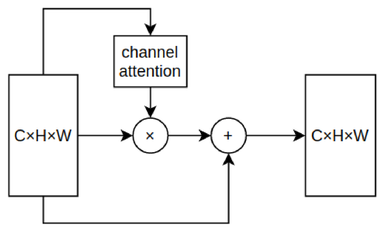

Figure 1:Illustration of our attention module. Here x denotes matrix multiplication and + denotes element wise sum. C, W and H respectively denotes channel, width and height of a layer respectively.

Results

Model	Frame(fps)	mIoU(%)
DPN (Yu et al., 2018b)	1.2	60.1
DeepLab (Chen et al., 2017)	4.9	61.6
ENet (Paszke et al., 2016)	-	51.3
ICNet (Zhao et al., 2018)	27.8	67.1
BiSeNet1 (Yu et al., 2018a)	-	65.6
BiSeNet2 (Yu et al., 2018a)	-	68.7
DFANet A (Li et al., 2019)	120	64.7
DFANet B (Li et al., 2019)	160	59.3
SwiftNet pyr (Orsic et al., 2019)	-	72.85
SwiftNet (Orsic et al., 2019)	-	73.86
Ours	124	74.12

Figure 3:Results using CamVid dataset. First column: input image from dataset, second column: predicted segmentation from our network and third column: ground truth segmentation.

Model	InputSize	FLOPs	Params	Time(ms)	Frame(fps)	mIoU(%)
PSPNet (Zhao et al., 2017)	713 × 713	412.2G	250.8M	1288	0.78	81.2
DeepLab (Chen et al., 2017)	512 × 1024	457.8G	262.1M	4000	0.25	63.1
SegNet (Badrinarayanan et al., 2017)	640 × 360	286G	29.5M	16	16.7	57
ENet (Paszke et al., 2016)	640 × 360	3.8G	0.4M	7	135.4	57
CRF-RNN (Zheng et al., 2015)	512 × 1024	-	-	700	1.4	62.5
FCN-8S (Long et al., 2015)	512 × 1024	136.2G	-	500	2	63.1
FRRN (Pohlen et al., 2017)	512 × 1024	235G	-	469	0.25	71.8
ICNet (Zhao et al., 2018)	1024 × 2048	28.3G	26.5M	33	30.3	69.5
BiSeNet1 (Yu et al., 2018a)	768 × 1536	14.8G	5.8M	13	72.3	68.4
BiSeNet2 (Yu et al., 2018a)	768 × 1536	55.3G	49M	21	45.7	74.7
DFANet A (Li et al., 2019)	1024 × 1024	3.4G	7.8M	10	100	71.3
DFANet B (Li et al., 2019)	1024 × 1024	2.1G	4.8M	8	120	67.1
Ours	1024 × 1024	1.8G	5.5M	6	134	72.4

Figure 4:Accuracy and speed analysis on Cityscapes test dataset.

Network Architecture

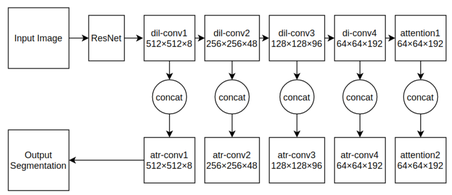

Figure 2:Illustration of our neural network architecture. Here dil-conv represents dilated convolutions and atr-conv represents atrous convolutions. attention1 and attention2 are the two channel wise attention modules used in this work.

Results

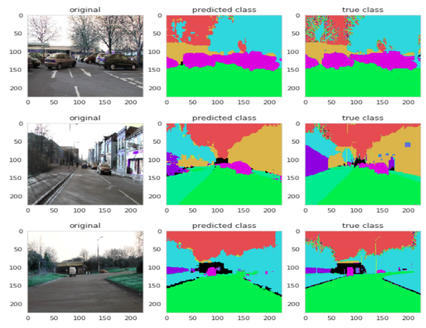

Figure 5:Results using CamVid dataset. First column: input image from dataset, second column: predicted segmentation from our network and third column: ground truth segmentation.

Conclusion

In this paper, we proposed a semantic segmentation network using multi scale attention feature maps and validated its performance on Camvid dataset and Cityscapes dataset. We used a downsampling and upsampling structure with dilated and atrous convolutional layers respectively with combinations between corresponding pooling and unpooling layers. We also propose our own attention module to enlarge the receptive field and encode more contextual information. Multi scale feature maps are merged using concat operator for encoding more contextual information. We present loss function, optimization details, ablation studies and evaluation metrics used. Our network achieves mean IOU value of 74.12 which is better than the previous state of the art on semantic segmentation while running at >100 FPS.

References

- C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 325–341, 2018a.
- C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1857–1866, 2018b.
- H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 7151–7160, 2018.

Contact Information

abhinavsagar4@gmail.com