Available online: February 5, 2020

Commun.Fac.Sci.Univ.Ank.Series A2-A3
Volume 62, Number 1, Pages 26–34 (2020)
DOI: 10.33769/aupse.611958
ISSN 1303-6009 E-ISSN 2618-6462
http://communications.science.ankara.edu.tr/index.php?series=A2A3

LEARNING DENSE CONTEXTUAL FEATURES FOR SEMANTIC SEGMENTATION

LONG ANG LIM AND HACER YALIM KELES

Abstract. Semantic segmentation, which is one of the key problems in computer vision, has been applied in various application domains such as autonomous driving, robot navigation, or medical imagery, to name a few. Recently, deep learning, especially deep neural networks, have shown significant performance improvements over conventional semantic segmentation methods. In this paper, we present a novel encoder-decoder type deep neural network-based method, namely XSeNet, that can be trained end-to-end in a supervised manner. We adapt ResNet-50 layers as the encoder and design a cascaded decoder that is composed of the stack of the X-Modules, which enables the network to learn dense contextual information and have wider field-of-view. We evaluate our method using CamVid dataset, and experimental results reveal that our method can segment most part of the scene accurately and even outperforms previous state-of-the art methods.

1. Introduction

Semantic segmentation, which is the key problem in the field of computer vision that concerned with partitioning image frames in video sequences into multiple regions, is an active research area until these days. Consider road-scene video sequences, where those road-scene regions are assigned any semantic categories such as sidewalk, road, pedestrian, sky, building, and tree etc. More precisely, semantic segmentation is a multi-class classification problem, where each pixel of an image is associated with a class label. In this case, the representation of an image is changed to something that is more meaningful and easier to understand. Pixel-wise semantic segmentation is useful in various applications such as autonomous driving, medical imagery, and robot navigation etc.

Recently, Convolutional Neural Networks (CNNs) [14] are very powerful in extracting hidden feature representations from data [28] and have been successfully

Received by the editors: August 27, 2019; Accepted: February 04, 2020.

Key words and phrases. Semantic segmentation, deep learning, convolutional neural networks, pixel classification, autonomous driving.

used in many recognition tasks [14, 13, 22, 16, 12, 2, 21]. In particular, Fully Convolutional Networks (FCNs) that are based on transfer learning [16] have shown significant performance improvement over traditional computer vision approaches by large margins. In this case, the knowledge that is gained from image classification problem is adapted to dense spatial class prediction domain where each pixel in an image is marked with a class label. However, due to feature resolution reduction caused by consecutive pooling and strided convolution operations in the pre-trained models, this is a trade-off for the segmentation problem since the contextual details or some object boundaries are lost. To recover lost information, some methods [21, 2] used a skip-connection technique to injecting decoder features by using encoder features. This technique enables the network to learning and refining lost object boundaries by using low level features in a sense of previous knowledge can be incorporated with prior knowledge to help boosting the network learning.

Motivated by the recent success of deep neural networks, we propose a simple yet effective encoder-decoder type network that can be trained end-to-end in a supervised manner. The rest of this paper is organized as follows: a brief summary about previous works is described in Section 2, our method is described in Section 3, our segmentation results are described in Section 4, and conclusion in Section 5.

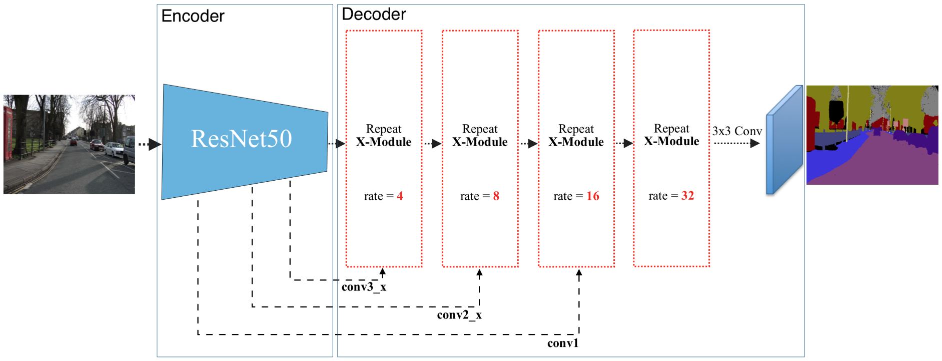

FIGURE 1. The flow of XSeNet architecture.

2. Related Works

Motivated by significant improvement of deep neural networks, most researchers explore the capabilities of such networks in semantic segmentation domain which make this domain improved over times. Recently, the FCNs was proposed by authors in [16] in semantic segmentation domain. The authors cast fully connected layers of the pre-trained network; i.e. AlexNet [13], VGG-Net [22], and GoogLeNet [23] into fully convolutional forms. The in-network upsampling and the skip architecture are introduced to refine the semantic outputs. Their method improves the

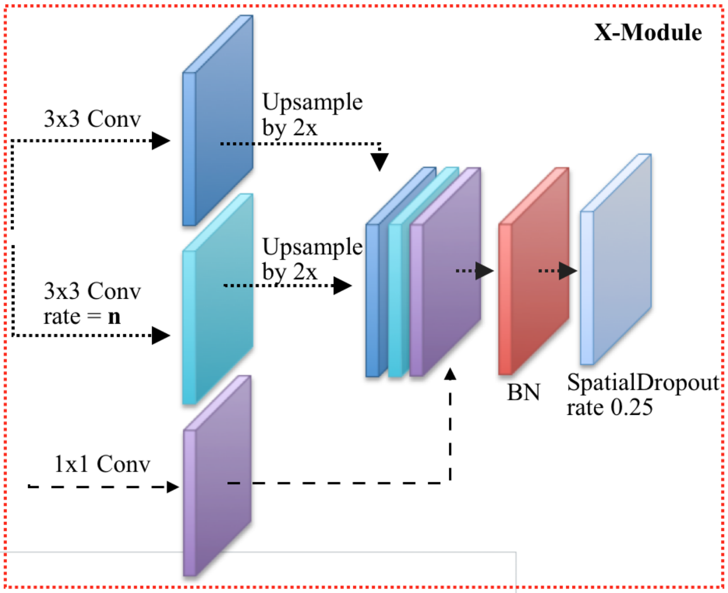

FIGURE 2. The X-Module

performance on several datasets such as Pascal VOC 2011-12 [7], SIFT Flow [15] etc. This work is considered as a milestone since the network can be trained end-to-end by taking input images of arbitrary sizes and producing dense predictions. Apart from this, a decoder approach is proposed by authors in [17] to mitigate the limitation of FCNs. In this work, the deconvolution layers and unpooling layers are embedded on top of the pre-trained classification model (VGG-16 Net) in a symmetric way. The network is applied to each object proposal in an input image, and then the resultant semantic outputs from all proposals were combined to construct the final segmentation output. This technique alleviates the limitation caused by the fixed receptive field of FCNs.

SegNet [2] is another state-of-the-art method and it is trained with CamVid dataset [3]. The network is composed of a decoder network embedded on top of the pre-trained VGG-Net. The decoder network consists of convolution layers and upsampling layers, where the max-pooling indices of the encoder network are used to perform non-linear upsampling in the decoder part. This network outperforms existing methods such as FCN [16] and DeconvNet [17].

Another recent encoder-decoder network is proposed by authors in [11]. In this work, the network is designed based on DenseNet [9] and trained from scratch without using any pre-trained networks. The proposed approach improves the state-of-the-art on CamVid and Gatech dataset [20]. Another efficient semantic segmentation method called ENet was proposed by [18]. The authors adopt the ideas of ResNet network [8] and specifically design the model for fast inferences and having low computational cost; yet, provides comparable results to existing methods. Authors in [1] proposed a residual coalesced convolutional network (RCC-Net) for video semantic segmentation problem. This encoder-decoder typed network architecture is designed based on the inspiration of ResNet model and Inception network [23]. For computational efficiency, the authors utilize an initial module to reducing the dimensionality of the input images into small feature maps before passing to the encoder network. Experimental results show that this network outperforms most of state-of-the-art methods. Authors in [25] proposed an encoder-decoder type recurrent neural network-based method to exploit contextual information from images and it provides promising results on various benchmarks.

Recently, authors in [19] proposed a two-branch network architecture which can segment high resolution images at 123.5fps that can be suited to efficient computation on embedded devices with low memory.

Moreover, dilated convolution or atrous convolution has been successfully used in semantic segmentation task [26, 5, 4, 6]. The idea of dilated convolution is to enlarge field-of-views in the network without learning extra parameters. To exploit this nice property, we also adopt this idea in our implementation.

3. Our Method

3.1. **Network Configuration.**

3.1.1. *The Encoder.* We adapt the pre-trained ResNet-50 [8], also known as ResNet 50 layers, and design a network architecture as depicted in Fig. 1 and Fig. 2. We utilize the first four blocks of ResNet-50 (conv1, conv2_x, conv3_x, conv4_x) and freeze all layers of the network, except the last identity block of conv4_x where we keep it for fine-tuning (for detailed ResNet network architecture, one may refer to the original paper). Skip connection technique has been successfully applied in image recognition task [8] to facilitate gradient flows through the network, and in semantic segmentation task [21] to allow the decoder to learn relevant features that are lost by pooling operations in the encoder part. Motivated by these ideas, we apply the skip layers in our implementation, and we found out that it makes the network converge faster and improves the accuracy.

3.1.2. *The Decoder.* Our decoder network (Fig. 2) contains four stacked modules, namely X-Module, where each module contains exactly the same configuration, except that the dilation rate is increased by a factor of two from the first module to the last module. For each module, given a set of feature map F, two convolutional operations are operated in parallel:

(1) Firstly, a fixed receptive field 3x3-convolution with a stride of 1 is operated on a given feature map F to produce a set of feature maps M of size $H \times W \times 64$. Then the feature map M is upsampled by a factor of 2 to produce a set of feature maps M of size $H' \times W' \times 64$.

(2) Secondly, a dilated 3x3-convolution with a dilation rate n and a stride of 1 is operated on the same feature map F to produce a set of feature maps N of size $H \times W \times 64$. Follow the same process, the feature map N is upsampled by a factor of 2 to produce a set of feature maps N of size $H' \times W' \times 64$.

(3) Thirdly, for the first, second and third module, a point-wise 1x1-convolution projects a set of high dimensional feature map depths in the encoder part to low dimension P of size $H' \times W' \times 64$. Note that the skip weight-layers in conv2_x and conv3_x of the encoder part are taken from the last identity block (we pick the weight-layers right before BatchNormalization [10] and addition operation, from this identity block). There is no skip layer (dash-line layer) in the fourth module.

(4) Finally, feature map M, N and P are concatenated along the depth axis to produce $H' \times W' \times 192$ feature maps (since $64 \times 3 = 192$), except the fourth module where there is no skip layer P is used. Note that the concatenated features are sequentially followed by BatchNormalization, ReLU, and SpatialDropout [24] layer. Then, the resultant feature maps are passed through the next layer.

Further Analysis: a fixed receptive field of normal 3x3-convolution can be interpreted as learning local features without incorporating global information into account, where dilated 3x3-convolution enables the network to have wider receptive fields that is significant for dense prediction by aggregating wide-ranged contextual information into account. In our implementation, these types of convolutional results are concatenated with previous knowledge in the encoder part, and then passed through the next layer to learning dense prediction. In each module (Fig. 2), since the concatenated feature map is a result of two convolutional operations that are operated on the same input, this concatenated feature map is expected to be strongly correlated. In this case, to alleviate overfitting, we apply the SpatialDropout right after this concatenated feature to drop correlated feature maps with a dropout rate of 25% and we found it to be effective in our implementation.

3.2. Training Details.

In this work, we utilize CamVid dataset [3], which consists of 367 training, 101 validation, and 233 testing images at 360x480 resolution. The CamVid challenge is to segment scenes into 11 classes such as road, building, car, pedestrian, sidewalk, traffic sign etc. We train our network using the same training frames as described in [2].

We perform data augmentations to expanding the training sizes by horizontally flipping, rotating +10/-10 degrees. To make the network paying attention to the rare classes such as pole or traffic light, we perform zooming and cropping (to left and right) on the input image by focusing on those rare classes. Note that we zoom the images to left and right in ratios of 100% and 200%, and then crop those images. To help the network paying attention to the rare class, we weight the classes differently by computing the class weights from entire training examples.

We train our network using Adam optimizer with a batch size of 1, setting the learning rate to 1e-3 and training by 15 epochs. We reduce the learning rate by a factor of 5 when the minimum validation loss stops improving for 5 epochs. Early stopping is applied when the minimum validation loss stops improving (min_delta equals to 1e-4) for 10 epochs. A softmax cross entropy loss is used in our training. Note that we resize the training images to a resolution of 352x480 in this experiment.

4. Results and Discussion

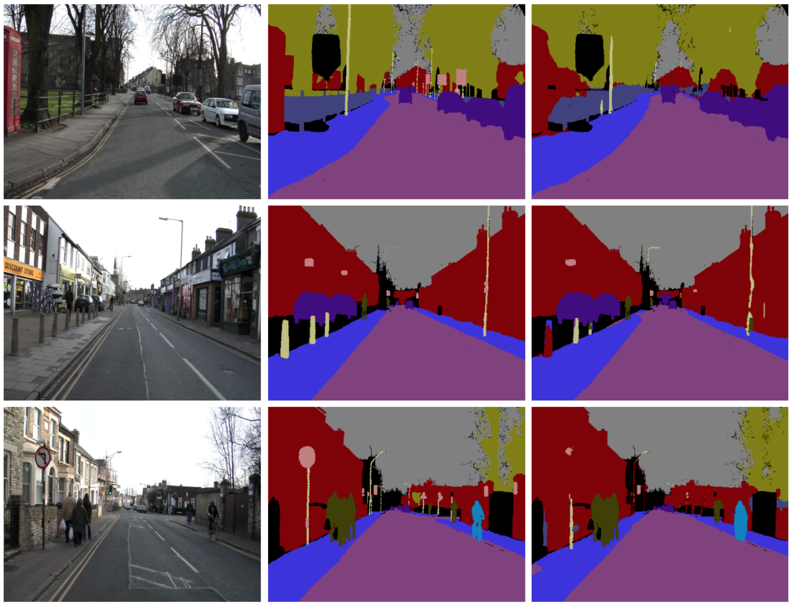

Figure 3. Our test results on CamVid dataset. First, second, and third column show input images, ground-truths, and our test results, respectively.

We measure the performance using three metrics followed the work in [2]; i.e. percentage of pixels correctly classified (G), mean predicted accuracy over all classes or class average accuracy (C), and mean intersection over all classes (mIoU).

Our test results and comparisons are depicted in Table 1. As can be seen, our method outperforms all listed methods by some margins. We also depict our segmentation results in Fig. 3 and Fig. 4. Most parts of the scenes are correctly classified by our method, especially major classes such as road, building, tree or car

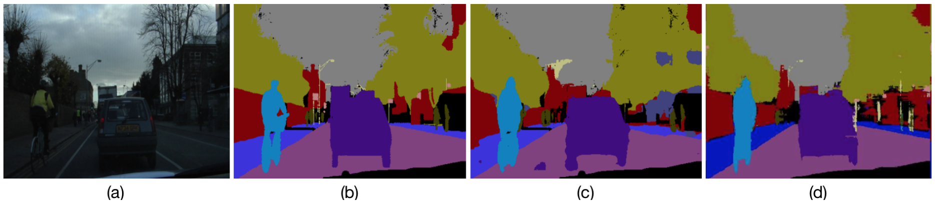

Figure 4. A comparison with the top model listed in Table 1. First, second, third, and fourth column show input images, groundtruths, our test results, and ReSeg results [25], respectively.

etc. One remarkable thing is that it is capable of distinguishing between bicyclists and pedestrians (Fig. 3, last row), where it is the failed case of other methods. This improvement may cause by aggregating contextual information using dilated convolution. However, it fails to detect rare classes, i.e. traffic lights and poles, where they share similar color intensities with the scene and it may be caused by the imbalanced data problem. Similarly, in the challenging scenario (dark scene), our method produces comparable results compared to the top listed model (Fig. 4).

Table 1. The *test results* on CamVid dataset and comparisons with the state-of-the-art methods. (G: Percentage of pixels correctly classified, C: Mean predicted accuracy over all classes, mIoU: Mean intersection over all classes.)

Methods	G	C	m-IoU
XSeNet (ours)	90.15	74.08	62.98
ReSeg [25]	88.70	68.10	58.80
RCC-Net [1]	–	71.50	53.30
ENet [18]	–	68.30	51.30
SegNet [2]	88.60	65.90	50.20
DeconvNet [17]	85.60	–	48.90
SegNet-Basic [2]	84.20	56.50	47.70
FCN [16]	83.50	57.30	47.00

5. Conclusion

In this research, we propose a novel network architecture and apply some useful techniques that significantly improve the segmentation results. We included four stacked X-Modules, all of which have exactly the same architectural configurations, except for the increased dilation rates. This configuration increases the receptive field size and helps better aggregating the contextual information without creating additional burden to the architecture. Our method produces good results and outperforms previous state-of-the-art methods in all metrics. We believe that our method can be improved further by increasing the layers in our network, i.e. ResNet-101 and ResNet-152 layers, or by replacing the pre-trained architecture (ResNet-50) with dilated residual networks [27], since this network is designed to alleviate the problem of losing relevant information when applying pooling operations. We will investigate the optimal solution for the aforementioned problems in our future research.

References

1. Igi Ardiyanto and Teguh Bharata Adji, *Deep residual coalesced convolutional network for efficient semantic road segmentation*, IPSJ Transactions on Computer Vision and Applications **9** (2017), no. 1, 6.
2. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*, arXiv preprint arXiv:1511.00561 (2015).
3. Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla, *Semantic object classes in video: A high-definition ground truth database*, Pattern Recognition Letters **30** (2009), no. 2, 88–97.
4. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*, IEEE transactions on pattern analysis and machine intelligence **40** (2018), no. 4, 834–848.
5. Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, *Rethinking atrous convolution for semantic image segmentation*, arXiv preprint arXiv:1706.05587 (2017).
6. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, *Encoder-decoder with atrous separable convolution for semantic image segmentation*, arXiv preprint arXiv:1802.02611 (2018).
7. Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, *The pascal visual object classes (voc) challenge*, International journal of computer vision **88** (2010), no. 2, 303–338.
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
9. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, *Densely connected convolutional networks.*, CVPR, vol. 1, 2017, p. 3.
10. Sergey Ioffe and Christian Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, arXiv preprint arXiv:1502.03167 (2015).
11. Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio, *The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation*, Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE, 2017, pp. 1175–1183.
12. Andrej Karpathy and Li Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.
13. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012, pp. 1097–1105.
14. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE **86** (1998), no. 11, 2278–2324.

15. Ce Liu, Jenny Yuen, and Antonio Torralba, *Sift flow: Dense correspondence across scenes and its applications*, IEEE transactions on pattern analysis and machine intelligence **33** (2011), no. 5, 978–994.
16. Jonathan Long, Evan Shelhamer, and Trevor Darrell, *Fully convolutional networks for semantic segmentation*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
17. Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, *Learning deconvolution network for semantic segmentation*, Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520–1528.
18. Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello, *Enet: A deep neural network architecture for real-time semantic segmentation*, arXiv preprint arXiv:1606.02147 (2016).
19. Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla, *Fast-scnn: fast semantic segmentation network*, arXiv preprint arXiv:1902.04502 (2019).
20. S Hussain Raza, Matthias Grundmann, and Irfan Essa, *Geometric context from videos*, Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 3081–3088.
21. Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-net: Convolutional networks for biomedical image segmentation*, International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
22. Karen Simonyan and Andrew Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556 (2014).
23. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, *Going deeper with convolutions*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
24. Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler, *Efficient object localization using convolutional networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648–656.
25. Francesco Visin, Marco Ciccone, Adriana Romero, Kyle Kastner, Kyunghyun Cho, Yoshua Bengio, Matteo Matteucci, and Aaron Courville, *Reseg: A recurrent neural network-based model for semantic segmentation*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.
26. Fisher Yu and Vladlen Koltun, *Multi-scale context aggregation by dilated convolutions*, arXiv preprint arXiv:1511.07122 (2015).
27. Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser, *Dilated residual networks.*, CVPR, vol. 2, 2017, p. 3.
28. Matthew D Zeiler and Rob Fergus, *Visualizing and understanding convolutional networks*, European conference on computer vision, Springer, 2014, pp. 818–833.

Current address: Long Ang Lim: Ankara University, Department of Computer Engineering, Ankara TURKEY

E-mail address: lim.longang@gmail.com

Current address: Hacer Yalim Keles (Corresponding author): Ankara University, Department of Computer Engineering, Ankara, TURKEY

E-mail address: hkeles@ankara.edu.tr

ORCID: http://orcid.org/0000-0002-1671-4126