Three-branch Semantic Segmentation Network

Lihua Wei and Yingdong Ma
School of computer science, Inner Mongolia University, Hohhot, 010021, China.
Email: 31709023@mail.imu.edu.cn, csmyd@imu.edu.cn

Abstract. Effective Semantic segmentation requires both spatial details and object-level semantic information. Meanwhile, context information is also important for complex scene understanding. However, it is hard to meet these demands simultaneously in the top-down CNN structure. In this paper, we tackle this problem with a Three-branch Semantic Segmentation Network (TSS Net). The proposed TSS Net consists of three parts, including a spatial network, a semantic network and a context network. The spatial network utilizes convolutional layers with small stride and a spatial pyramid pooling module to extract multi-scale spatial features. In the semantic network, multiple level features are combined to enhance semantic information. The context network integrates different scales contextual information to facilitate objects localization in complex scene. The proposed semantic segmentation framework has been evaluated on the CamVid and the Cityscapes datasets. Experimental results demonstrate that the TSS Net achieves state-of-the-art performance.

1. Introduction

Semantic segmentation is one of the fundamental tasks in computer vision, which assigns labels to each pixel of input images. Most computer vision applications, including automatic driving, indoor navigation, human-computer interaction, and virtual reality, rely on precise semantic segmentation results. With the fast development of convolutional neural networks (CNNs), impressive results have been reported in semantic segmentation due to the network ability of automatic learning of hierarchical features [1, 2, 3, 4].

Recent works on semantic segmentation shows that there are mainly three approaches to combine spatial features with high-level semantic information. Firstly, the FCN [1] series replaces the fully connected layers with the convolutional layers to form a full convolutional network. Object and scene segmentation are conducted from deep layer feature maps which can be enhanced by fusing previous layer features. Nevertheless, prediction based on high-level feature maps may lead to lower accuracy. Secondly, image pyramid-based methods [3] transform input images into different scales and apply CNN model on each level images to generate multiple level features. Predictions from different levels are combined to obtain the final output. Although these methods are simple, they might lose spatial details in the downsampling process and increase computational cost. Thirdly, some segmentation works adopt the Encoder-Decoder structure [2]. These methods use low-level features to restore spatial resolution in the decoder stage. However, some detail structure information might be lost in the downsampling stage and only part of them can be recovered in the upsampling stage.

Context relationship has been proved to be important especially for complex scene understanding. This problem can be remedied by using the contextual information. In [4], global context is obtained from global average pooling. Although global pooling is commonly used in various visual applications, global descriptors are not suitable for semantic segmentation due to loss of spatial relationship. The works in [4, 5] extracts contextual information by using dilated convolutions. Dilated convolutions play an important role in maintaining spatial resolution of deep feature maps.
Inspired by these works, we propose a three-branch semantic segmentation network to alleviate these problems. The proposed TSS Net consists of a spatial network, a semantic network and a context network. The semantic network is designed to enhance multiple level semantic information, while the spatial network focuses on learning spatial details from high resolution feature maps. The context network combines multiple scale features to capture local and global context information.

2. Related Work

2.1. Contextual and Spatial Information
Semantic segmentation requires context information to make reliable prediction in complex scene and classification of objects with similar appearance. The pyramid scene parsing network [4] aggregates scene context features from different sub-regions which strength network capability of global scene understanding. In [7], Yang et al. proposed the DenseASPP which connects a set of atrous convolutional layers to obtain dense feature maps for autonomous driving.

2.2. Attention Mechanism
Recently, attention mechanism has been successfully applied in many computer vision tasks, such as image classification [8], and image segmentation [5,9]. Hu et al. proposed a Squeeze and Excitation structure [8] to adjust output response by modeling the relationship between channel features. In the Pyramid attention network [9], Li et al. combines attention mechanism with pyramid structure.

2.3. Multi-branch Networks
One of the main limitations of a top-down CNN architecture is that high resolution features obtained from shallow layers are lack of high-level semantic meaning. To alleviate the problem, some semantic segmentation approaches adopt the multi-branch architecture. The Bisenet [10] employs a spatial path to preserve spatial information for generating high-resolution features. Meanwhile, a context path with fast downsampling is used to yield sufficient receptive fields. The ContextNet [11] adopts a deep branch captures global context information with a shallow branch that focuses on spatial details.

3. Approach

3.1. Overview

![Figure 1. The overall structure of the Three-branch Semantic Segmentation Network. FEM: Feature Enhancement Module. CPAM: Context Pooling Aggregation Module. PPM: Pyramid Pooling Module](image)

In the semantic segmentation, both spatial details and object-level semantic information are crucial. In addition, local and global contextual information is also necessary. In this work, we introduce a three-branch semantic segmentation network. The proposed TSS Net consists of three main components: the
spatial network, the semantic network and the context network. The spatial network obtains spatial details. The semantic network captures semantic information and the context network collects multiple scale context information. The overall framework is shown in Figure 1.

3.2. Spatial Network
We propose the spatial network to enhance fine-level features. The spatial network consists of three convolution layers to preserve spatial details. The batch normalization [12] and the rectified linear unit (ReLU) [13] are utilized for normalization and activation. An attention module is appended to each convolution layer, which is designed to adjust output response for regions of interesting. These layers are followed by a Pyramid Pooling Module (PPM) to integrate multi-scale spatial information.

Pyramid Pooling Module: As shown in Figure 2, the PPM fuses features from four scales. Let input feature maps have size of \( W \times H \times D \), where \( D \) is the number of channels. Features in each level have size \( \frac{W}{k} \times \frac{H}{k} \), \( k = [1, 2, 3, 6] \). We use a \( 1 \times 1 \) convolution to reduce channels to \( D/4 \). The feature maps with smaller sizes are upsampled to \( W \times H \). Features of different pyramid scales are then concatenated with input feature maps as the output features which have spatial size of \( W \times H \times 2D \).

Attention Module: An attention module is applied to learn weight, as illustrated in Figure 3. Input features of the attention module contain the output feature maps of the feature enhancement module and the spatial network convolution layer. These features are concatenated and followed by a \( 3 \times 3 \) convolution. The global pooling is utilized to generate global context. The attention module makes feature refinement by learning feature weight to improve segmentation accuracy.

3.3. Semantic Network
In the proposed method, the ResNet [14] is adopted as backbone in the semantic network. The semantic network has two purposes. Firstly, the FEM is introduced to combine multiple level features and provide enhanced features to the spatial network. Secondly, feature maps of different blocks are concatenated to yield deep feature block and shallow feature block for the context network.

Feature Enhancement Module: As shown in Figure 4, features of two levels, \( F_n \) and \( F_{n+1} \) are concatenated to obtain combined feature \( F \). Feature \( F \) is transformed by a \( 3 \times 3 \) convolution to obtain feature \( F' \). After reducing feature channels by using a \( 1 \times 1 \) convolution, dilated convolutions are utilized in parallel to expand receptive field. The dilation rate determines size of the receptive field and helps to aggregate context information. The reduction ratio controls computational overheads. According to the experience, we set \( \{ d = 4, r = 16 \} \).

3.4. Context Network
As shown in Figure 5, the deep feature block and the shallow feature block are combined to form input features of the CPAM module. We use deconvolution to unify the size of features of different blocks. Let the \( W \times H \) input feature maps have \( D \) channels. The pyramid pooling fuses feature maps under two pyramid levels with kernel size of \( 2 \times 2 \) and \( 4 \times 4 \), respectively. Then we upsample the smaller size features to get the same size features as the original feature maps. Finally, two level features are concatenated with the input features to get the output features with spatial size of \( W \times H \times 2D \). Combination of different pooling layer features not only expands the receptive field, but also lead to better utilization of multi-scale contextual information.
4. Experiments

4.1. Implementation Details
In this work, the ResNet-50 [14] is adopted as the baseline model. It’s parameters are learned on the ILSVRC 2012 dataset [15]. Our implementation is based on the platform TensorFlow using the Root Mean Square prop (RMSprop) optimization. The initial learning rate is 0.0001 and the decay is 0.995. The input images are cropped to patches with the size of 512×512 when training on the Camvid [16]. The images are 768×768 when training on the Cityscapes [17]. We use the mean intersection-over-union (Mean IoU) and class average accuracy (ClassAvg) to evaluate performance.

4.2. Ablation Study
Feature Enhancement Module: Table 1 shows the results of using FEM modules. Experimental results demonstrate that utilization of FEM improves performance significantly. Compared to the baseline, using of three FEM modules increases Mean IoU from 72.71% to 78.45%.

Attention Modules: Results of experiments with or without attention modules (Att) are listed in the first three rows of Table 1. The experiment achieves 80.76% Mean IoU. These experiments demonstrate that the attention module selects important spatial information.

Multi-Scale Pyramid Pooling Module: As shown in Table 1, the segmentation performance increased to 81.80% by using the proposed PPM module. The experiment proves that collecting multiple scales spatial context information is necessary for semantic segmentation applications.

Context Pooling Aggregation Module: The context pooling aggregation module is implemented for encoding different scales context information. The Table 1 shows that utilizing CPAM module increases performance from 81.80% to 82.90%.

Table 1. Ablation study of different structure on CamVid val dataset.

| baseline | FEM | Att | PPM | CPAM | MeanIoU(%) | ClassAvg(%) | Parameters(M) |
|----------|-----|-----|-----|------|------------|-------------|---------------|
| √        |     |     |     |      | 72.71      | 87.38       | 89.81         |
| √        | √   |     |     |      | 78.45      | 90.21       | 164.50        |
| √        | √   | √   |     |      | 80.76      | 90.34       | 170.84        |
| √        | √   | √   | √   |      | 81.80      | 90.47       | 175.84        |
| √        | √   | √   | √   | √    | 82.90      | 90.55       | 183.87        |

4.3. Experimental Results on CamVid
We compare performance of the proposed approach with state-of-the-arts on the CamVid dataset. The results are shown in Table 2. Our method outperforms most state-of-the-arts. We obtain 82.2% and 83.1% Mean IoU. Figure 6 shows that most performance improvements come from accurate location of small objects as some small parts, such as traffic sign, pole and bicyclist.
(a)Ground Truth  (b)Baseline  (c)PSPNet  (d)TSS Net

Figure 6. Visualization results on CamVid test set.

Table 2. Performance comparison with state-of-the-art methods on CamVid test dataset.

| Method          | Building | Tree | Sky  | Car  | Sign | Road | Pedestrian | Fence | Pole | Sidewalk | Bicycles | Mean IoU(%) |
|-----------------|----------|------|------|------|------|------|------------|-------|------|----------|----------|------------|
| FCN8[1]         | 77.8     | 71.0 | 88.7 | 76.1 | 32.7 | 91.2 | 41.7       | 24.4  | 19.9 | 72.7     | 31.0     | 57.0       |
| Dilation8[4]    | 82.6     | 76.2 | 89.0 | 84.0 | 46.9 | 92.2 | 56.3       | 35.8  | 23.4 | 75.3     | 55.5     | 65.3       |
| VideoGCRF[18]   | 86.1     | 78.3 | 91.2 | 92.2 | 63.7 | 96.4 | 67.3       | 63.0  | 34.4 | 87.8     | 66.4     | 75.2       |
| ours (ResNet50) | 87.2     | 87.2 | 94.8 | 87.6 | 68.6 | 93.2 | 68.9       | 85.3  | 58.8 | 87.8     | 84.8     | 82.2       |
| ours (ResNet101)| 91.8     | 84.6 | 97.8 | 91.7 | 76.5 | 91.5 | 66.2       | 83.6  | 60.3 | 83.1     | 87.3     | 83.1       |

4.4. Experimental Results on Cityscapes
Performance comparison with some recent segmentation works on the Cityscapes test dataset are shown in Table 3. The proposed TSS Net achieves 81.3% Mean IoU. The performance improvement is due to the multi-branch network structure, which utilizes different modules, to learn representative multi-level features for semantic segmentation.

Table 3. Performance comparison on Cityscapes test set.

| Methods         | Backbone   | Mean IoU |
|-----------------|------------|----------|
| BiSeNet[10]     | ResNet101  | 78.9     |
| DFN[5]          | ResNet101  | 79.3     |
| DenseASPP[7]    | DenseNet161| 80.6     |
| TSS Net         | ResNet101  | 81.3     |

5. Conclusion
In this paper, we present a three-branch network for semantic segmentation. The multiple level features are integrated in the semantic network branch to extract high-level semantic information. The spatial details are enhanced in the spatial network branch by aggregating spatial features from different regions. The context information is collected in the context network to further improve segmentation accuracy. The feature enhancement module and the context pooling aggregation module are also introduced to enhance network representative capability and embed multiple level context information. Experimental results on the CamVid and Cityscapes datasets demonstrate superior semantic segmentation performance of the new framework as compared with state-of-the-arts.

6. Acknowledgements
This work was supported by the National Natural Science Foundation of China under Grant 61461039.
7. References

[1] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[2] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[3] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[4] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[5] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1857-1866).

[6] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[7] Yang, M., Yu, K., Zhang, C., Li, Z., & Yang, K. (2018). Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3684-3692).

[8] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[9] Li, H., Xiong, P., An, J., & Wang, L. (2018). Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.

[10] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 325-341).

[11] Poudel, R. P., Bonde, U., Liwicki, S., & Zach, C. (2018). Contextnet: Exploring context and detail for semantic segmentation in real-time. arXiv preprint arXiv:1805.04554.

[12] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[13] Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323).

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016, October). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630-645). Springer, Cham.

[15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

[16] Brostow, G. J., Shotton, J., Fauqueur, J., & Cipolla, R. (2008, October). Segmentation and recognition using feature from motion point clouds. In European conference on computer vision (pp. 44-57). Springer, Berlin, Heidelberg.

[17] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

[18] Chandra, S., Couprie, C., & Kokkinos, I. (2018). Deep spatio-temporal random fields for efficient video segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8915-8924).