Aggregation Architecture and All-To-One Network for Real-Time Semantic Segmentation

Kuntao Cao, Xi Huang, Jie Shao

University of Electronic Science and Technology of China, Chengdu, 611731, China

{caokuntao, xihuang}@std.uestc.edu.cn and shaojie@uestc.edu.cn

Abstract

Deep convolutional neural network has demonstrated its outstanding performance in the field of image semantic segmentation. However, the enormous computational complexity of existing high-precision networks limits the application of the model in real-time segmentation tasks. How to achieve a good trade-off between accuracy and speed becomes a challenge. Existing solutions can be roughly divided into three categories according to the network architecture: dilation, encoder-decoder, and multi-pathway, each of which has its advantages.

Contributions

In this paper, we make the following contributions:

a) First, unlike the previous three architectures, we propose a new aggregation architecture as the network backbone.

b) Second, a multi-level auxiliary loss design model is used for the training phase, which can improve the model segmentation effect.

c) According to this aggregation structure, an all-to-one network (ATONet) for real-time semantic segmentation is proposed, which achieves a good trade-off between speed and accuracy by assembling the features of all blocks.

The proposed network achieves the accuracy of 74.4% and 70.1% mIoU with the inference speed of 42.7 FPS and 93.5 FPS on the Cityscapes and CamVid.

Introduction

Semantic segmentation, which aims to assign semantic labels to each pixel in the image, is a fundamental but challenging task in computer vision. It has a number of potential applications in the field of autonomous driving, medical image diagnostics, video surveillance, indoor scene understanding, and so on.

Models with huge computational costs are difficult to meet the real-time requirements of applications such as autonomous driving. To solve this problem, many researchers have proposed some lightweight convolutional network structures to explore how to reduce the number of network calculations while ensuring a certain degree of accuracy. These methods can be divided into three main architectures: (1) dilation architecture, replacing traditional downsampling operations with dilated convolution to generate high resolution and semantically informative features, as shown in Figure 1 (a); (2) encoder-decoder architecture utilizing top-down and skip connection to reuse low-level high-resolution features, as illustrated in Figure 1 (b); (3) multi-pathway architecture, which integrates multiple pathways that focus on different features to ensure the segmentation effect, and the core lies in branch design and fusion structure design (see Figure 1 (c)).

Methods

Aggregation Architecture. The proposed aggregation architecture is inspired by the encoder-decoder structure and the multi-pathway structure, hoping to use a single branch to obtain more semantic information while retaining the characteristic spatial information. Unlike the way that the encoder-decoder structure merges layer by layer, the aggregation structure tends to directly sum the features of each layer. This structure integrates the output of the features by each module in the network backbone into the same dimension, and then sums them, as shown in Figure 1(d).

Multi-level Auxiliary Loss. The multi-level auxiliary loss strategy is only used in the training phase of the network, which means that the strategy will not affect the inference speed of the network during the prediction process. Besides, unlike the previous method of calculating the auxiliary loss through up-sampling features, we use downsampling (the sampling strategy is nearest neighbor interpolation) groundtruth to calculate the auxiliary loss, with the purpose of making the network pay more attention to the main loss rather than the auxiliary loss.

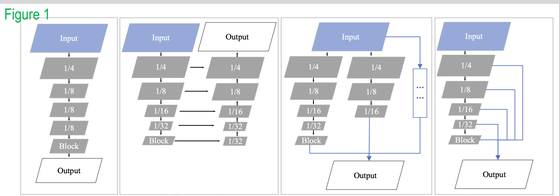
$$w_i = \begin{cases} 1, & if\ use\ loss_i \\ 0, & if\ not\ use\ loss_i \end{cases}$$

$$sum\_loss = main\_loss + \sum_{i=1}^{5}(w_i \times loss_i)$$

All-to-One Network. According to the network design of the aggregation structure, ATONet is proposed. As shown in Figure 2, we use ResNet-18 as the backbone network and further utilize SPP modules with pooling sizes of 8, 4, and 2 to expand the receptive field. The output features of each module of the backbone are upsampled to a quarter of the original input image through the up module and the number of feature channels is processed into 128 dimensions by 1 x 1 convolutional layer. The `final' block consists of a standard 3 x 3 convolution, BN, and ReLU, which is a more common combination. A 3 x 3 convolution, BN, ReLU, dropout, and 1 x 1 convolution constitute the final seg module, which processes the number of feature channels into the number of segmentation classes. Dropout is used to avoid overfitting, and its probability of an element to be zeroed is set to 0.1. The proposed multi-level auxiliary loss strategy is used to improve the segmentation performance. The total loss can be calculated as

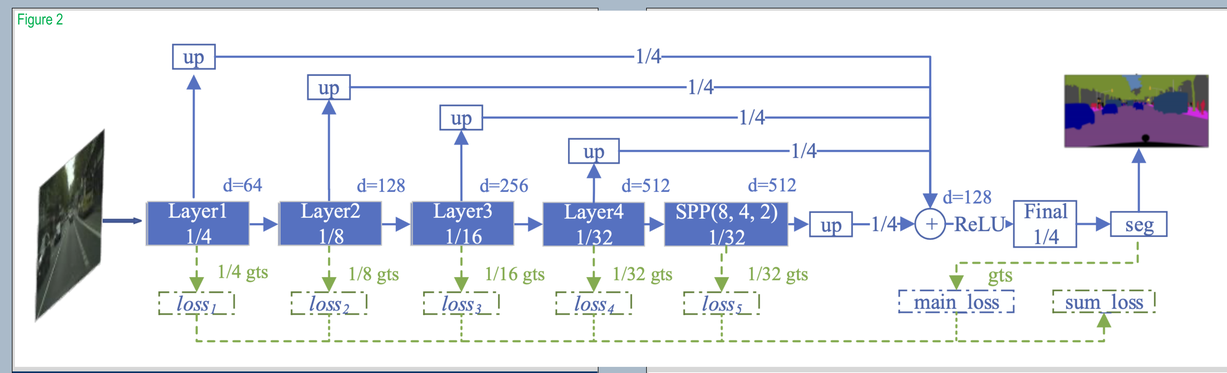
$$sum\_loss = main\_loss + loss_3 + loss_4 + loss_5$$

Figure 1



(a) Dilation architecture (b) Encoder-decoder architecture (c) Multi-pathway architecture (d) Aggregation architecture

Figure 2



Experiments

Table 1. Comparisons of the proposed method and other state-of-the-art methods on the Cityscapes test dataset. '-' indicates that the corresponding result is not provided by the method. For SwiftNetRN-18 [19], label 'ens' means the ensemble of single scale model and pyramid model, and label 'pyr' demotes the pyramid fusion model.

Method	Input size	FLOPs (G)	Params. (M)	FPS	mIoU (%)
SegNet [11]	360 × 640	286	29.5	16.7	57
ENet [9]	360 × 640	3.8	0.4	135.4	57
FCN-8s [1]	512 × 1024	136.2	-	2.0	63.1
FRRN [16]	512 × 1024	235	-	2.1	71.8
ERFNet [13]	512 × 1024	27.7	2.1	41.7	69.7
TwoColumn [17]	512 × 1024	57.2	-	14.7	72.9
BiSeNet [18]	768 × 1536	14.8	5.8	105.8	68.4
DFANet A [14]	1024 × 1024	3.4	7.8	100	71.3
DFANet B [14]	1024 × 1024	2.1	4.8	120	67.1
DFANet A' [14]	512 × 1024	1.7	7.8	160	70.3
SQNet [12]	1024 × 2048	270	-	16.7	59.8
ICNet [15]	1024 × 2048	28.3	26.5	30.3	70.6
SwiftNetRN-18 ens [19]	1024 × 2048	218.0	24.7	18.4	76.5
SwiftNetRN-18 pyr [19]	1024 × 2048	114.0	12.9	34.0	75.1
SwiftNetRN-18 [19]	1024 × 2048	104.0	11.8	39.9	75.5
ATONet	1024 × 2048	134.5	13.3	42.7	74.4

Table 2. Comparisons on the CamVid test dataset. The resolution of the test data is 720×960. '-' indicates that the corresponding result is not provided by the method.

Method	FPS	mIoU (%)
SegNet [11]	4.6	60.1
DeepLab [3]	4.9	61.6
ENet [9]	61.2	51.3
ICNet [15]	27.8	67.1
SwiftNetRN-18 pyr [19]	-	73.86
BiSeNet [18]	175	65.6
DFANet A [14]	120	64.7
DFANet B [14]	160	59.3
ATONet	93.5	70.1

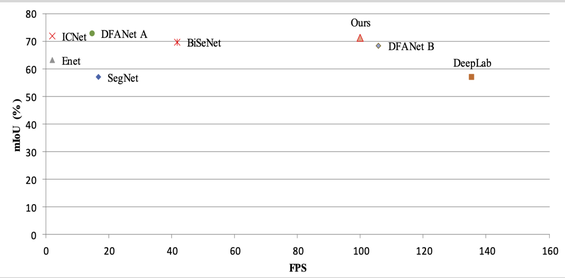


Figure 3 Comparison of the test results on the Camvid data set with scatter plots.

Table 3. Speed and accuracy comparison of aggregation architecture against other architectures on the Cityscapes validation dataset. All networks are trained for 100 epochs.

Architecture	Resolution	FLOPs (G)	Params. (M)	FPS	mIoU (%)
Encoder-decoder	1024×1024	62.02	12.49	31.0	67.39
Multi-pathway	1024×1024	87.42	12.67	23.1	67.82
Aggregation	1024×1024	58.05	12.58	39.0	68.41

Table 4. Ablations for SPP pooling size and multi-level auxiliary loss on the Cityscapes validation dataset. † indicates that the feature is upsampled to groundtruth in the auxiliary loss calculation.

Epochs	Val. resolution	SPP pooling size	Aux $loss_1$	Aux $loss_2$	Aux $loss_3$	Aux $loss_4$	Aux $loss_5$	mIoU (%)
100	1024×1024	(6, 3, 2, 1)						67.79
	1024×1024	(8, 4, 2, 1)						68.03
	1024×1024	(8, 4, 2)						69.02
100	1024×1024	(8, 4, 2)	✔					67.54
	1024×1024	(8, 4, 2)	✔	✔				66.68
	1024×1024	(8, 4, 2)	✔	✔	✔			66.70
	1024×1024	(8, 4, 2)	✔	✔	✔	✔		68.62
	1024×1024	(8, 4, 2)	✔	✔	✔	✔	✔	69.01
	1024×1024	(8, 4, 2)		✔	✔	✔	✔	69.55
	1024×1024	(8, 4, 2)			✔	✔	✔	70.21
	1024×1024	(8, 4, 2)				✔	✔	69.70
	1024×1024	(8, 4, 2)					✔	68.84
100	1024×1024†	(8, 4, 2)			✔	✔	✔	69.52
100	1024×2048	(8, 4, 2)			✔	✔	✔	74.21
200	1024×2048	(8, 4, 2)			✔	✔	✔	75.58

Conclusion

In this paper, a new real-time general-purpose semantic segmentation network architecture, aggregation architecture is proposed. Besides, a multi-level auxiliary loss model is used for the training phase. To further verify the effectiveness of our proposed aggregation, an all-to-one network (ATONet) is designed. Finally, our experiments verify the effectiveness of the proposed methods. In the future, we plan to extend our work in several aspects, such as redesigning a more lightweight backbone network to replace ResNet-18, conducting more detailed multi-level auxiliary loss combination experiments, and changing the loss of each module weight coefficient. The code of ATONet is publicly available at https://github.com/KTMomo/ATONet.