LIP Net: Real-Time Semantic Segmentation of Person Body Parts

LIP Net: Real-Time Semantic Segmentation of Person Body Parts

Talha Ilyas[1], Hyongsuk Kim[2*]
[1)] 전북대학교, 전자 정보공학부(TEL: ※ 개인정보 표시제한 E-mail: talha@jbnu.ac.kr)
[2)] 전북대학교, 전자 정보공학부(TEL: 063-270-2477; E-mail: hskim@jbnu.ac.kr)

Abstract Human visual understanding and pose estimation in wild scenarios is one of the fundamental tasks of computer vision. Traditional deep convolution networks (DCN) use pooling or subsampling layers to increase the receptive field and to gather larger contextual information for better segmenting human body parts. But these subsampling layers reduce the localization accuracy of the DCN. In this work, we propose a novel DCN, which uses artuous convolution with different dilation rates to probe the incoming feature maps for gathering multi-scale context. We further combine a gating mechanism which recalibrates the convolutional feature responses adaptively by learning the channel-wise statistics. This gating mechanism helps to regulate the flow of salient features to the next stages of network. Hence our architecture can focus on different granularity from local salient regions to global semantic regions, with minimum parameter budget. Our proposed architecture achieves a processing speed of 49 frames per second on standard resolution images.

Keywords semantic segmentation, encoder-decoder architecture, pose estimation, human parsing

1. Introduction

Human body-part parsing is a challenging task in computer vision due to the articulation of body limbs, self-occlusion and diverse clothing styles. Significant improvements have been achieved by DCN. However, for cluttered background with objects which are similar to body-parts or with heavy occlusion body-parts, DCN face difficulty to classify and detect each body parts correctly.

Since the dawn of Fully Convolutional Networks (FCNs) [1] semantic segmentation has gained a lot of popularity and following the main idea of embedding low contextual information in a progressive manner to preserve spatial and temporal information a lot of encoder-decoder architecture have been introduced in literature. SegNets [2] introduced unpooling (i.e. inverse of pooling) to upsample the score maps in a gradual way. To remedy the loss of localization information by subsequent downsampling of feature maps U-net [3] proposed skip-connections between encoder and decoder to preserve spatial information. Further, the intermediate layers were exploited by RefineNet [4] with the skip-connections, which uses multipath refinement via different convolutional modules to get final predictions. PSP-Net [5] used spatial pyramid pooling at different scales and Deeplab [6] used artuous convolutions with different dilation rate for exploiting multi scale information. More recently, networks like Dilated ResNet (DRN) [7] used artuous convolutions to increase the valid receptive field. Furthermore, Deeplab-v3+ [8] combined dilated convolution with depthwise separable convolution . By doing so they achieved significant performance boost. Using multi-scale contextual information has proven essential for vision tasks. Intuitively, larger context region captures global spatial configurations of object, while smaller context region focuses on the local part appearance.

Yet, from a practical perspective, many applications cannot fully enjoy the high accuracy demonstrated by recent advances. The reason is for this is the bulk of current work assumes the abundance of computational resources (e.g. GPUs, memory, power) to run these models which for many applications are not available. Besides being challenging, the problem of human parsing under low memory and computing capacity has received little attention from the research community so far. In contrast to previous approaches, we propose a novel real-time segmentation network for human body-part parsing. In our network we use the dilated separable convolution for capturing global and local context in conjunction with a gating mechanism for modelling channel-wise dependencies. We verify the effectiveness of our model on a benchmark dataset namely Look into Person (LIP).

2. Proposed Approach

2.1 Network Architecture

※ 이 논문은 2020 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.NRF-2019R1A6A1A09031717)또한, 이 논문은 농촌진흥청 “농업과학 기술 개발 협력 연구 프로그램(No. PJ01389105)” 의 지원으로 수행된 연구임.

A modified FCN for real-time human body-part parsing, namely LIP-Net is shown in Figure 1.

Network backbone consists of six stages. Among those, first four stages $S_{i\in\{1,2,3,4\}}$ are followed by subsequent pooling operations for reducing feature map size. In the next stages $S_{i\in\{5,6\}}$ we don't perform pooling operation. Because, after using four subsequent pooling operations in first four stages the extracted feature map size is 16 times smaller than the input at the end of encoder. Reducing it further will result in loss of a lot of useful localization information, making the decoding process more difficult. In the first two stages the dilation rate is set to d=1, and in the next three stages the dilation rate is doubled for every next stage i.e. $d_{i\in\{3,4,5\}}$= {2,4,8} for stages $S_i\in\{3,4,5\}$. The final stage S6 again has a dilation rate of d6 =1 to avoid the gridding artifact [59].

All stages consist of two SE-ResNet (SER) [9] like blocks. At each stage inside each SER instead of using simple convolution, we decided to use the dilated separable convolution, which is a combination of dilated and depth-wise separable convolution. It allows the network designer to freely control the feature map's size and filter's effective receptive field (ERF), while significantly reducing the network computational cost. This decomposition allows the DCNN to achieve better performance with much less parameters. In dilated convolutions, ERF can be easily changed by changing the dilation rate 'd_i', where normal convolution being a special case of dilated convolution with d=1. Increasing the ERF at each stage of network helps the convolutional filters to aggregate multi-scale contextual information more efficiently. For details about SER-modules we refer interested readers to [83].

As we go deeper into the DCNNs, even though the deeper layers have a large theoretical receptive field (TRF) but their effective receptive field (ERF) is much smaller than the theoretical one. Information regarding global context plays a vital role in scene segmentation. So, at the end of the encoder we probe the feature maps of the last stage (i.e. S_6) for aggregating global and sub-region context by incorporating ASPP [8] module shown in Figure 1. ASPP act as a hierarchal global prior using dilated convolution at different dilation rates to extract global contextual information from S_6's feature maps at multiple scales.

Our decoder bilinearly upsmaples the feature map by a factor of 16 in subsequent steps. In first step the output of ASPP is bilinearly upasmpled by a factor of 2 and concatenated with the feature maps of 3rd stage of encoder. The incoming feature maps of 3rd stage are first reduced by a factor of four by 1x1 convolution, so that they don't overweigh the feature enriched feature maps of ASPP. These concatenated feature maps are then processed through a 3x3 convolution and are upsampled 2nd time by a factor of 4. Lastly the network's output is obtained performing a 1×1 convolution followed by Softmax activation on the upsampled feature maps of the 2nd step.

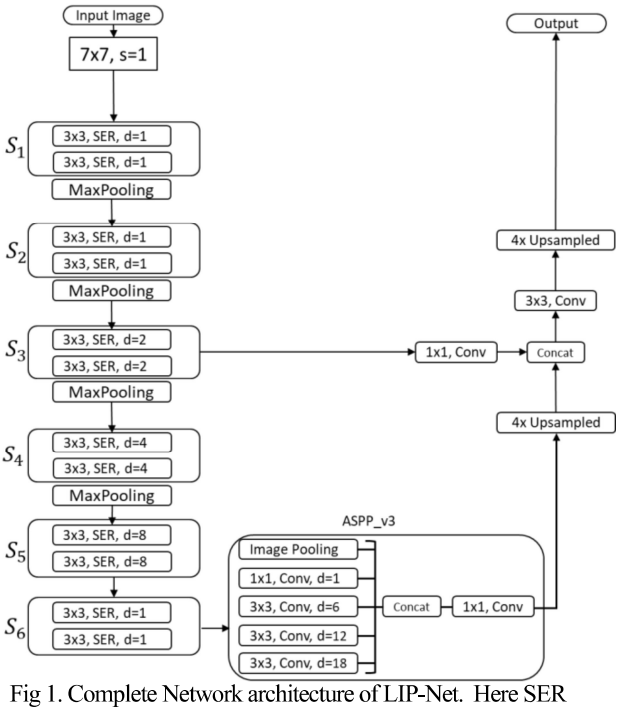

Fig 1. Complete Network architecture of LIP-Net. Here SER represents SE-ResNet Block, 's' and 'd' are stride and dilation rate respectively. S_i represents the different stages of network, where i∈{1,2,3,6}.

3. Experiments

3.1 Implementation Details

Firstly, we resized all the images and segmentation masks to 512×512 resolution, for reducing training time and memory requirements. In encoder for number of channels in each stage, we set C=16. At each stage, the number of channels (C) and the dilation rate (d) are successively increased as shown in Figure 1. For SER block, (Figure 1) following [83] we set the reduction ratio r=8. In ASPP module for global context aggregation, following [8] we set dilation rate d={1,6,12,18}, respectively for three parallel branches as shown in Figure 1. For training, following we employ Adam optimizer along with poly learning rate policy where, $l_{r\_new} = l_r * \left(1 - \frac{iter}{total\_iter}\right)^{power}$ .Here we set power=0.9 and l_r=0.005. For loss function we used weighted cross entropy. We used dropout of 0.25 and set the mini batch_size = 4. The network is trained for 20K iterations

Method	Param. (M)	mIoU (%)	FPS (sec)
FCN-8s	134.3	59.7	2
Seg-Net	29.46	64.8	16.7
Deeplab_v2	48.0	68.7	2.4
LIP-Net	**3.97**	**70.3**	**49**

Table 1. Experimental results on LIP Dataset.

3.2 LIP Dataset

LIP data set is a large-scale dataset which focus on semantic understanding of a person's body. Which contains the high-level annotation of 50K images.

Annotations belong to 19 semantic human body parts and various clothing as shown in Figure 2. The training, validation and testing splits contain 30K, 10K and 10K images respectively.

3.3 Results and Discussion

The human part segmentation results on LIP dataset are reported in Table 1. The models are compared based on three benchmark metrics number of parameters, mean intersection over union (mIoU) and processing speed (FPS). As our goal is to make a real-time segmentation network so, we focus mainly on FPS and number of parameters of the networks. LIP-Net achieves 1.6% more mIoU than the Deeplab_v2 but significantly outperforms it in processing speed. It can be clearly seen from the Table 1 that only LIP-Net crosses the real-time performance barrier of 30fps.

Fig 2. Qualitative results on LIP dataset

4 Conclusion

In this paper we proposed an architecture for real-time pixel-wise semantic segmentation. In contrast to high accuracy state-of-the-art approaches that assumes the abundance of computational resources to be available to run these models. The goal of this paper was to perform human body parsing from a practical perspective, because many applications cannot fully enjoy the high accuracy demonstrated by recent computationally heavy models. Our proposed algorithm captures local and global context, controls the flow of salient features in the successive stages of a network via a gating mechanism while still maintaining real-time performance. Our future work will include human pose estimation and experiments on power consumption and compression technique of a model.

References

[1] E. Shelhamer, J. Long, and T. J. I. A. o. t. H. o. C. Darrell, "Fully convolutional networks for semantic segmentation," no. 04, pp. 640-651, 2017.

[2] V. Badrinarayanan, A. Kendall, R. J. I. t. o. p. a. Cipolla, and m. intelligence, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," vol. 39, no. 12, pp. 2481-2495, 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015: Springer, pp. 234-241.

[4] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925-1934.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. J. I. t. o. p. a. Yuille, and m. intelligence, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," vol. 40, no. 4, pp. 834-848, 2017.

[7] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 472-480.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801-818.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.