MPFANet: Semantic Segmentation Using Multiple Path Feature Aggregation

Wang-Su Jeon¹ and Sang-Yong Rhee²

¹Department of IT Convergence Engineering, Kyungnam University, Changwon, Korea

²Department of Computer Engineering, Kyungnam University, Changwon, Korea

Ijfis

Abstract

Image segmentation is the process of simplifying the analysis of the meaning or the front to say the process of dividing the image into a set of multiple pixels. The multiple path feature aggregation (MPFA) method proposed in this paper aims to extract various information of an object, and uses conventional pyramid pooling or the extraction of various sized features. This information can be combined with different regional features to obtain the overall feature information. We split four paths to extract numerous local features, and the results showed that the mean intersection over union (mIOU) is 81.6% for the validation data from the PASCAL VOC 2012 dataset, and a better performance than the existing DeepLab model was demonstrated.

Keywords: MPFA, Semantic segmentation, Feature aggregation, CNN, Inverted residual block, Local context

1. Introduction

The performance of robots and self-driving cars is improving owing to the recently developed hardware, big data technology, and artificial intelligence. Requires the robot and the autonomous navigation technology is basically a visual recognition, and visual recognition is extracted feature data in a meaningful image entered through the camera detects an object, it refers to the ability to recognize, classify.

For the visual perception, which is basically categorized into the foreground and background, and in the case of traditional computer vision, image segmentation was performed through image segmentation. Image segmentation refers to the process of dividing an image into several sets of pixels. The purpose of segmentation is also a process of simplifying a scene based on its meaning or interpretation, rather than simply classifying it. In particular, a boundary or set with an object is determined.

Image segmentation can also be divided into three types, as shown in Figure 1 [1], instance segmentation, semantic segmentation, and panoptic segmentation [2]. Instance segmentation refers to classifying each object possessed by a pixel, and semantic segmentation refers to assigning a pixel value for each class. In addition, panoptic segmentation refers to a method of classifying the foreground as an instance segmentation and the background as a semantic segmentation and classifying as “stuff” or “things.” The term stuff refers to countless backgrounds, and things refers to countable objects.

The technology of existing computer vision after the use of deep learning has not achieved

Received: Nov. 4, 2020
 Revised : Sep. 23, 2021
 Accepted: Nov. 8, 2021

Correspondence to: Sang-Yong Rhee
 (syrrhee@kyungnam.ac.kr)
 *These authors contributed equally to this work.
 ©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure 1. Type of segmentation: (a) input, (b) instance segmentation, (c) semantic segmentation, and (d) panoptic segmentation. The images are from [1].

a good performance in the field of image segmentation can see what looks outstanding performance [3,4], which has also included Cityscapes [5], PASCAL VOC [6], and COCO [7] in recent years. Through data provided by competition, it can be seen that the performance has improved significantly.

When applying a convolutional neural network (CNN) [8–10], three problems occur: a reduction in the resolution, a loss of information, objects, and space invariance. These problems are the first, occurs in the pooling process, the second, is generated by the existence of an object between the layers and the layer or between the extracted feature map. Third, it is an object is literally learning the spatial information of the entire image less affected due to the surrounding environment when you want to present in the center there is a disadvantage vulnerable to rotate or noise. Although this can be solved using techniques such as data augmentation, doing so creates another problem that requires a long learning time.

In this paper, we propose a multiple path feature aggregation (MPFA) method to solve this problem. This method can improve the performance by using learning to create a feature map and spatial information in a variety of sizes.

The remainder of this paper is organized as follows: Section 2 discusses related work, and Section 3 describes how the proposed model applies MPFA. Section 4 details the experimental results and analysis of the proposed method, and Section 5 provides some concluding remarks and areas of future research.

2. Related Work

Semantic segmentation can use the models [11–14] in this category that have been proven in terms of performance based on four challenge images as a backbone, and provides a better performance through fine-tuning. A fully convolutional network is the starting point of its improved semantic segmentation performance by changing the existing fully connected approach to a fully convolution approach [3]. In addition, the method of increasing the receptive field in viewing neural network objects [15–17] is a method that uses a dilated convolution or deconvolution [18]. To carry out the semantic segmentation, DeepLab [16] and PSPNet [19], which are the most frequently used models, employ multi-scale segmentation.

When the size of the reduced or increased layer is characterized by a loss of information, i.e., in this case, when the correction information of the previous higher-level layer is used, the loss can be reduced. Therefore, a multi-scale is applied because it takes a method of complementary features of different sizes [20–24] and can improve the performance, increasing the class and location information for image segmentation. Additionally, a conditional random field (CRF) is used post-processing to improve the performance of the predicted result.

By learning to end-to-end, not to apply the CRF with post-processing methods but to improve the network, improve [25–27] and the location information of an object, the objects overlap or complex scenes have a poor performance. Therefore, to better take advantage of the features of various sizes by using non-traditional methods [28,29] through deep learning, this method has an improved performance in detecting objects [30]. Liu et al. [31] found that it can be used to improve global average pooling (GAP) with the fully convolutional network (FCN).

It is, therefore, possible to use the information in the global spatial features through the GAP information and combining different characteristic maps for local information through a pyramid network.

3. MPFA

The backbone network of the model proposed in this research is a fine-tuning of the ResNet-50 layer model, and MPFA consists of four paths, and three of these paths have four sub-paths. As shown in Figure 2, a feature map of size $41 \times 41 \times 512$ is made by concatenating the outputs of the three paths and the ones of the backbone. Next, after adjusting the size to 21 by

Figure 2. Multiple path feature aggregation network architecture.

Figure 3. Atrous spatial pyramid pooling structure.

Figure 4. Multi-scale residual feature structure.

1×1 convolution to predict the class, training and prediction are performed.

3.1 Multi-Scale Feature

The proposed model used for the first path is the atrous spatial pyramid pooling (ASPP) of DeepLab, as shown in Figure 3. ASPP may be in the number of computations while maintaining the size and area of the receptive field of various sizes. The atrous rate used in this study apply to $\{6,8,16,24\}$.

3.2 Multi-Scale Residual Feature

The second path of the proposed model is composed of a layer that performs the residual operation in a pyramidal format, as shown in Figure 4. The overall structure uses the inverted residual block shown in Figure 5(a), and this method is used in MobileNet v2 [32]. Inverted residual block, so that the dimensions are expressed in that advantage that it is possible

Figure 5. Linear bottleneck structure: (a) residual block and (b) inverted residual block.

MobileNet v2 and then expanded to a higher order channel expressed in the low-order channel by the high-level information through the hypothesis of a manifold to minimize the loss of information through ReLu.

3.3 Pyramid Pooling Module

The third path of the proposed model is the module used in PSP-Net. In the case of average pooling used here, after obtaining the average of the channels, partial information of a $1/4$ size

Figure 6. Pyramid pooling module structure.

is extracted through the operation of each pooling, and then connected. At this time, the parameters of the average pooling are set to 7, 9, 11, and 15, respectively, and the structure is shown in Figure 6.

4. Experiments

4.1 Experiment Details

The experiment environment used in this study was as follows. The operating system used an Intel Xeon Gold 5120 CPU in an Ubuntu 18.04.4 environment, 32 GB of memory, and RTX TITAN X 24GB graphics cards. The deep learning framework uses TensorFlow 1.14.0, and PASCAL VOC 2012 was used for learning. The PASCAL VOC 2012 video segmentation dataset consists of 10,582 pieces of training data and 1,449 pieces of test data, with a total of 20 classes. The model learning applied 20,000 iterations, a batch size of 8, a learning rate of 0.00001, and the stochastic gradient descent method was used for Adam. Then, by using the constructed model, the performance change of the model was measured by experimenting while controlling the number of iterations according to the number of MPFAs and batch size of the proposed module.

4.2 Experiment Result

The experiment results obtained using the proposed method are shown in Figures 7 and 8. The input image, ground truth, DeepLab [16] result, and the proposed model result is shown in order from the left. In the case of Figure 7, bottles, persons, buses, horses, and TVs are classified. When comparing DeepLab and the proposed model, it can be seen that it is similar to deeplab or misclassifier error than DeepLab. This problem is judged to be a classifier error owing to the large amount of information on the scale. Figure 8 shows the classification of people, bottles, horses, and cows. In Figure 8, it can be seen that the overall partitioning performance is excellent. In the

Table 1. Mean IOU and fps measurement according to MPFA size

Method	Mean IOU(%)	fps
ResNet50-MPFA512	81.6	12
ResNet50-MPFA256	78.3	24
ResNet50-MPFA128	75.5	32
ResNet50-MPFA64	72.3	40

Table 2. Mean IOU measurement according to batch size and iterations

Method	Batch size	Iteration	Mean IOU(%)
ResNet50-MPFA512	4	10k	77.0
ResNet50-MPFA512	4	20k	77.7
ResNet50-MPFA512	10	20k	78.3
ResNet50-MPFA512	10	20k	81.6

Table 3. PASCAL-VOC-2012 validation data of each model compared with mean IOU

Method	Mean IOU(%)
ResNet-GCN [36]	81.0
DFN [35]	80.6
DeepLab v3+ [34]	79.3
FastDenseNas-arch0 [33]	78.0
DeepLab-CRF [16]	77.69
Proposed method	81.6

second image, one can see that there is a part that recognizes the error of a horse identified as a cow.

The results of the evaluation and comparison of the overall performance are shown in Table 1. Table 1 shows the results of measuring the performance change by adjusting the feature size of the proposed method. For the performance, the accuracy (mean intersection over union [mIOU]) and frames per second (fps) are shown. When the feature size is the largest, it can be seen that the performance is good, and when the feature size is 256, there is little deterioration in performance, and 32 frames can be calculated.

As shown in Table 2, the performance was compared by adjusting the batch size and the number of iterations, and when the batch size was 10, the highest performance was shown. Table 3 is shown the comparison of the model in the PASCAL VOC 2012 Validation Leader Board that evaluate the performance as a valid data DeepLab-CRF [16], FastDenseNas-arch0 [33], DeepLab v3+ [34], DFN [35], it represents a ResNet-GCN [36]

Figure 7. Visual measurement of PASCAL VOC 2012 data. Results of the proposed method are compared with the baseline.

with the result of comparing the five model and performance.

5. Conclusion

In this paper, we proposed an MPFA method based on PASCAL VOC 2012 data and the Deeplab model to conduct semantic

segmentation. The performance results of the proposed method show an accuracy of 81.6% coteuna, demonstrating a better performance than the Deeplab model class. There was a problem in extracting the non-labeled regions and a degraded classification performance occurred. Therefore, if there is a lot of scale information of the model, the performance will be poor and the

Figure 8. Visual measurement of PASCAL VOC 2012 data. The results of the proposed method are compared with the baseline.

approach is judged to be unsuitable.

In future research, we plan to develop a lightweight model suitable for real-time use while improving the classification performance and improving it into a model suitable for real-time application.

Conflict of Interest

No potential conflicts of interest relevant to this article are reported.

Acknowledgement

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1F1A1075968).

References

- [1] D. Mechea, "What is Panoptic Segmentation and why you should care," 2019 [Online]. Available: <https://medium.com/@danielmechea/whatis-panoptic-segmentation-and-why-you-should-care-7f6c953d2a6a>.
- [2] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA 2019, pp. 9404-9413.
- [3] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: integrated recognition, localization and detection using convolutional networks," 2013 [Online]. Available: <https://arxiv.org/abs/1312.6229>.
- [4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.
- [5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 3213-3223.
- [6] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 891-898.
- [7] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48
- [8] H. W. Choe, S. M. Park, and K. B. Sim, "CNN-based speech emotion recognition using transfer learning," *Journal of Korean Institute of Intelligent Systems*, vol. 29, no. 5, pp. 339-344, 2019. <https://doi.org/10.5391/JKIS.2019.29.5.339>
- [9] W. Y. Lee, K. E. Ko, Z. W. Geem, and K. B. Sim, "Method that determining the Hyperparameter of CNN using HS algorithm," *Journal of Korean Institute of Intelligent Systems*, vol. 27, no. 1, pp. 22-28, 2017. <https://doi.org/10.5391/JKIS.2017.27.1.022>
- [10] S. Kim and Y. Cho, "An artificial intelligence Othello game agent using CNN based MCTS and reinforcement learning," *Journal of Korean Institute of Intelligent Systems*, vol. 30, no.1, pp. 40-46, 2020. <https://doi.org/10.5391/JKIS.2020.30.1.40>
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.

- [12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014 [Online]. Available: <https://arxiv.org/abs/1409.1556>.
- [13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 1-9.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770-778.
- [15] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014 [Online]. Available: <https://arxiv.org/abs/1412.7062>.
- [16] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," 2016 [Online]. Available: <https://arxiv.org/abs/1606.00915>.
- [17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015 [Online]. Available: <https://arxiv.org/abs/1511.07122>.
- [18] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1520-1528.
- [19] 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*, Honolulu, HI, 2017, pp. 2881-2890.
- [20] F. Xia, P. Wang, L. C. Chen, and A. L. Yuille, "Zoom better to see clearer: human and object parsing with hierarchical auto-zoom net," in *Computer Vision – ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 648-663. https://doi.org/10.1007/978-3-319-46454-1_39
- [21] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 447-456.
- [22] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: deep feature aggregation for real-time semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019, pp. 9522-9531.
- [23] H. Choi, H. J. Lee, H. J. You, S. Y. Rhee, and W. S. Jeon, "Comparative analysis of generalized intersection over union and error matrix for vegetation cover classification assessment," *Sensors and Materials*, vol. 31, no. 11, pp. 3849-3858, 2019. <https://doi.org/10.18494/sam.2019.2584>
- [24] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 12, no. 4, article no. 701, 2020. <https://doi.org/10.3390/rs12040701>
- [25] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1377-1385.
- [26] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr, "Higher order conditional random fields in deep neural networks," in *Computer Vision – ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 524-540. https://doi.org/10.1007/978-3-319-46475-6_33
- [27] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 1529-1537.
- [28] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, 2016, pp. 2169-2178.
- [29] A. Lucchi, Y. Li, X. Boix, K. Smith, and P. Fua, "Are spatial and global constraints really necessary for segmentation?," in *Proceedings of 2011 International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 9-16.
- [30] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, "Scalable, high-quality object detection," 2014 [Online]. Available: <https://arxiv.org/abs/1412.1441>.

- [31] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: looking wider to see better," 2015 [Online]. Available: <https://arxiv.org/abs/1506.04579>.
- [32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 4510-4520.
- [33] V. Nekrasov, H. Chen, C. Shen, and I. Reid, "Fast neural architecture search of compact semantic segmentation models via auxiliary cells," 2019 [Online]. Available: <https://arxiv.org/abs/1810.10804>.
- [34] S. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. H. Torr, "Res2Net: a new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652-662, 2021. <https://doi.org/10.1109/TPAMI.2019.2938758>
- [35] 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*, Salt Lake City, UT, 2018, pp. 1857-1866.
- [36] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, pp. 4353-4361.

Wang-Su Jeon received his B.S. and M. S. degrees in Computer Engineering and IT Convergence Engineering from Kyungnam University, Masan, Korea, in 2016 and 2018, and is currently pursuing the Ph.D. degree in IT Convergence Engineering at Kyungnam University, Masan, Korea. His present interests include computer vision.
E-mail : jws2218@naver.com

Sang-Yong Rhee received his B.S. and M.S. degrees in Industrial Engineering from Korea University, Seoul, Korea, in 1982 and 1984, respectively, and his Ph.D. degree in Industrial Engineering at Pohang University, Pohang, Korea. He is currently a professor at the Computer Engineering, Kyungnam University, Masan, Korea. His research interests include computer vision, augmented reality, deep learning and human-robot interface.
E-mail : syrhee@kyungnam.ac.kr