Complex Document Layout Segmentation Based on An Encoder-Decoder Architecture

Jia Yao¹, Linlin Huang¹*

¹ School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, 100044, China
*Corresponding author’s e-mail: huangll@bjtu.edu.cn

Abstract. In our work, we propose an end-to-end encoder-decoder network for complex document layout segmentation. The proposed multi-scale feature extraction network with two parallel branches is applied to further process the feature maps, where one branch enriches the multi-scale information of the feature maps by building feature pyramids, another branch is introduced to capture the dependencies between different locations and integrate long-range context information. Moreover, we merge the outputs of the two branches to enhance the feature representation so as to further improve segmentation accuracy. Experimental results on datasets of PubLayNet and DSSE-200 demonstrate the effectiveness of our proposed method, which yields pixel-wise accuracy of above 99%.

1. Introduction
Document layout analysis and understanding technology has effectively promoted the digital management of information and widely used in daily life, such as postal automation and ancient manuscript recognition. Document layout segmentation studied in our paper is the most essential and critical step for all further processing including optical character recognition (OCR) and the goal is to split a document image into different regions: background, text, picture, list, etc. However, due to the flexible and complex layout, the changeable element shape, similar categories and the large difference in area size, the research of the document layout segmentation algorithm still faces huge challenges. In our work, we propose a method based on deep learning. Our main contributions are threefold.

- We use a new document image dataset named PubLayNet and achieve good performance.
- In order to improve the feature representation, we propose a new network named multi-scale feature extraction network (MFN) which consists of deconvolution pyramid pooling module (DPM) and position attention module (PAM) to process the feature map based on encoder-decoder structure. This is a novel network structure that is different from the previous one.
- The proposed MFN contains two parallel branches and the effective combination of the two branches makes the network fully utilize the global and local information of the image and improve the segmentation accuracy.

2. Related works
In 2017, Wick et al. proposed a high-performance network based on fully convolutional network (FCN)[1] for processing historical document segmentation, which directly learns from the original pixels, omitting the pre-processing steps[2]. Yang et al. proposed a more efficient end-to-end network that uses both the appearance information and semantic information to the segment document image[3]. However, the FCNs methods have a common shortcoming that some small objects are easily lost in
down-sampling stage. Ronneberger et al. use the symmetrical encoder-decoder architecture which can integrate the deep and shallow information through a simple skip-connection[4,5]. Many researchers think multi-scale strategy is conducive to integrating context information to solve the local ambiguity problem[6]. Consequently, Zhao et al. proposed the pyramid scene parsing network (PSPNet). This method of fusing local and global features together for classification prediction has significantly improved the segmentation accuracy[7]. Moreover, DeepLabs have proposed by Chen et al., the introduced dilated convolution has a larger receptive field but maintains the same amount of computation as ordinary convolution[8,9]. Then in order to obtain multi-scale context information, innovatively proposed the ASPP module, which uses multiple sampling rates to capture context information at multiple scales[10,11]. In 2019, Fu et al. proposed dual-channel attention network (DANet) to improve the classification accuracy by capturing rich contextual dependencies[12].

Combining the above methods and characteristics of document layout, we summarize the following points that need to be addressed when designing the network. Firstly, the shallow detail information should be kept as complete as possible in order to avoid the problem of small objects such as the bullets of list being ignored during down-sampling. Secondly, integrate richer long-range context information as much as possible to help the network better understand the complex layout and improve the accuracy of discrimination between similar categories. Thirdly, attention should be paid to obtaining multi-scale information, so that the performance for objects with large differences in size can be improved. Accordingly, we build our network from the above three aspects to further enhance the feature representation and achieve high-precision document segmentation.

3. Proposed method

we propose a new method following the symmetrical encoder-decoder structure. Specifically, inspired by PSPNet[7] and DANet[12], we design MFN which can integrate multi-scale information and global context information to further process the feature map. It is worth noting that different from previous document segmentation methods, any post-processing procedures which can further improve the model performance is not necessary in our network, but we still achieve better results.

3.1. Network structure

The framework of our proposed method is illustrated as figure 1. In encoder stage, we employ a pretrained ResNet101[13], the layers of Resnet101 correspond to enc-N(N=1,2,3,4,5) in our network. Note that we change the down-sampling stride of layer1 to 2 and introduce dilation convolution with dilation 2 in layer4. Then the feature map is sent to proposed MFN and we fuse the outputs of the two modules, specifically, we firstly concatenate the outputs of DPM and PAM to obtain a new feature map, and then we perform element-wise sum between it and the original feature map to obtain the final output of MFN. In decoding stage, we apply the bottleneck structure to build decoder network following resnet101, but we have fewer layers. In order to achieve pixel-wise segmentation, we perform upsampling to restore the feature map size. Furthermore, we adopt skip-connection to keep the shallow detailed information.

Figure 1. The overall structure of the proposed network. Segmentation colors are: text, table, title.
3.2. Multi-scale feature extraction network

MFN includes two branches DPM and PAM. We propose the DPM which extracts multi-scale features from different sub-regions by creating a feature pyramid. As shown in figure 2(a), the feature map highlighted in blue is extracted by ResNet101. In DPM, varying-size pooling kernels are chosen to perform pooling operations on the feature map, then the feature maps with varied sizes are sent to go through the 1×1 convolution layer to reduce dimension. Then we directly up-sample and concatenate these feature maps to form a thicker feature map as the output of DPM. DMP is based on PPM illustrated as figure 2(b) and has made certain improvements. The difference between them is that in PPM, bilinear interpolation is applied to up-sample feature maps, while in DPM we use deconvolution[14] instead, which through continuous learning for better parameters to restore the feature, so that the feature map contains richer semantic context information. In addition, compared with scene image, there are fewer categories in document, so considering segmentation efficiency, we modify the number of pyramid levels and use 3-level pyramid with bin size of 1 × 1, 2 × 2, 6 × 6. For the pooling operation, we choose average pool between max and average because there is no parameter to optimize in global average pooling thus overfitting is avoided in this layers.

**Figure 2.** The structure of the proposed DPM(a) and PPM(b) of PSPNet[7].

However, for pixel-wise segmentation, the direct connection between the underlying pixels is also crucial for classification. Therefore, another branch PAM is introduced next, which is proposed by DANet[12]. PAM focuses on modeling the relationship between pixels, enhancing the relationship between pixels with different positions but the same category. PAM selectively updates the features at each location via aggregating the features at all locations with weighted summation, the similar features have the same response regardless of the distance, thus improving intra-class compact and semantic consistency. As shown in figure 3. PAM uses matrix multiplication to model any two positions to reflect the similarity between them.

\[
S_{ji} = \frac{\exp(B_j \cdot C_i)}{\sum_{j=1}^{N} \exp(B_j \cdot C_j)}
\]

(1)

\[
E_j = \alpha \sum_{i=1}^{N} (S_{ji}D_j) + A_j
\]

(2)
Where \( s_{ij} \) represents the influence of the \( i^{th} \) position on the \( j^{th} \) position. If the two positions belong to the same category, then \( s_{ij} \) will produce a large value. \( \alpha \) is a variable scale factor, initialized as 0 and gradually learns to assign more weight. \( E \) is the final output of PAM.

4. Experiment

To evaluate the proposed method, we carry out a series of comprehensive experiments on PubLayNet[15] and DSSE-200[3] datasets, the experimental results prove the effectiveness of proposed method.

4.1. Dataset

PubLayNet is a large dataset for documents layout analysis proposed in 2019, containing over 360 thousand documents images. Typical document layout elements are assigned labels from the following dictionary: figure, paragraph, list and paragraph. DSSE-200 is a data set used in MFCN papers, including 200 images which involves 6 foreground region classes and one background class.

4.2. Experimental detail

Our end-to-end network is trained on four GTX 1080 Ti GPUs with mini-batch size of 8 and optimized by Stochastic Gradient Descent (SGD). We dynamically adjust the learning rate with iterations instead of fixing it. The basic learning rate and power is set to 0.005 and 0.9 respectively. So the current learning rate equals to the base one multiplying \( (1 - \frac{\text{iter}}{\text{max iter}})^{\text{power}} \). Momentum and weight decay are set to 0.9 and 0.0001 respectively. The performance has been improved by increasing the iteration number, which is set to 160K for PubLayNet experiment and 1.5K for DSSE-200. The whole implementations are conducted on PyTorch1.0.

4.3. Experimental results

In our work, we apply different feature extraction networks to conduct experiments on PubLayNet dataset. As the results shown in table 1, the feature extraction effect of ResNet101 is significantly better than VGG19[16], because the deep network structure helps to improve the feature representation. SegNet is better than FCN because its symmetrical structure is help to restore the feature, and PSPNet achieves the best performance which can attribute to PPM. Therefore, we use SegNet with ResNet101 as the basic model for subsequent experiments.

| network | backbone | mIoU    | accuracy |
|---------|----------|---------|----------|
| FCN     | VGG19    | 83.63%  | 94.17%   |
| UNet    | 4-layers | 86.51%  | 94.98%   |
| SegNet  | VGG19    | 87.34%  | 95.27%   |
| PSPNet  | ResNet101| 91.93%  | 96.35%   |

The MFN is a combination of two branch networks DPM and PAM, we need to evaluate the effectiveness of a single module and the combination, so experiments are carried on the PubLayNet dataset. From the results, the following conclusions can be drawn, we compare the mIoU of each method.
Firstly, as shown in table 2, we apply our proposed modules to replace the PPM, the results show that MFN brings 2.42% improvements compared with PSPNet, which proves the effectiveness of our method. Secondly, we apply PPM, DPM, PAM and MFN to the encoder-decoder structure respectively, results as shown in table 3. This novel structure with PPM achieves better result of 92.29%, which is 1.19% higher than the result of basic model. The results have proved the point of view in previous work, that is, the importance of Multi-scale context aggregation for segmentation task. We further evaluate the effectiveness of proposed module DPM, the result clearly shows that employing DPM outperforms the model with PPM by 1.99%, and brings 3.18% improvements compared the basic model. Then for PAM, we observe a mIoU of 92.69%, 1.59% better than the basic model, the improvements demonstrates that integrating richer global context information is the key to achieve high-precision document segmentation. Finally, the best performance is obtained when we apply MFN, the model yields a result of 95.42%, which brings 3.49% improvement compared with the PSPNet. The solid and consistent improvements demonstrate the effectiveness of the proposed method.

| Table 2. Evaluation results on PubLayNet dataset. DPM represents ResNet+DPM structure. |
|-------------------------------|------------|------------|------------|
| network | backbone | mIoU | accuracy |
| PSPNet | ResNet101 | 91.93% | 96.35% |
| DPM | ResNet101 | 93.14% | 97.69% |
| PAM | ResNet101 | 92.10% | 97.24% |
| MFN | ResNet101 | 94.35% | 98.72% |

| Table 3. Evaluation results on PubLayNet dataset. +DPM represents SegNet+DPM structure. |
|-------------------------------|------------|------------|------------|
| network | backbone | decoder | mIoU | accuracy |
| PSPNet | ResNet101 | -- | 91.93% | 96.35% |
| SegNet | VGG19 | VGG19 | 87.34% | 95.27% |
| SegNet | ResNet101 | B-decoder | 91.10% | 97.01% |
| +PPM | ResNet101 | B-decoder | 92.29% | 97.58% |
| +DPM | ResNet101 | B-decoder | 94.48% | 98.84% |
| +PAM | ResNet101 | B-decoder | 92.69% | 98.36% |
| +MFN | ResNet101 | B-decoder | 95.42% | 99.34% |

As shown in table 4, the comparison is conducted between our method and MFCN[3] on DSSE-200 dataset. The remarkable improvement of 4.5% indicates the effectiveness of proposed method.

| Table 4. Evaluation results on DSSE-200 dataset. |
|-------------------------------|-----------|------------|
| network | backbone | mIoU |
| MFCN | Dilated Block | 73.3% |
| ours | ResNet101 | 77.81% |

5. Conclusion
In this paper, we have proposed an encoder-decoder structure for complex document layout segmentation. Note that in the entire network, MFN is the key to effectively improve the feature representation. More specifically, the DPM structure focuses on extracting the multi-scale feature information, while the introduction of PAM is to integrate richer long-range context information from global. Experimental results on some public datasets show that the proposed method achieves good performance. Finally, we’d like to say that the problem of parameter compression is also a challenge we must solve in the further study to improve the speed of document layout segmentation.

Acknowledgments
This work has been supported by the Major Project for New Generation of AI under Grant No. 2020AAA0109702.
References

[1] Long, J., Shelhamer, E., Darrell, T. (2015) Fully convolutional networks for semantic segmentation. J. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651.

[2] Wick, C., Puppe, F. (2018) Fully Convolutional Neural Networks for Page Segmentation of Historical Document Images. In: 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 287–292.

[3] Yang, X., Yumer, E., Asente, P., Kraly, M., Kifer, D., Giles, C.L. (2017) Learning to extract semantic structure from documents using multimodal fully convolutional neural network. J. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4342-4351.

[4] Ronneberger, O., Fischer, P., Brox, T. (2015) U-net: convolutional networks for biomedical image segmentation. In: Cham: Springer Int. Publishing, pp. 234–241.

[5] Badrinarayanan, V., Kendall. (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. J. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): pp. 2481-2495.

[6] Garcia-Garcia A, Orts-Escolano S, Oprea S. (2017) A Review on Deep Learning Techniques Applied to Semantic Segmentation. J. Computer Science, ArXiv, volume:abs/1704.06857.

[7] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017) Pyramid scene parsing network. J. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 6230-6239.

[8] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. J. Computer Science, ArXiv, volume:abs/1706.05587.

[9] Yu, F., Koltun, V. (2016) Multi-scale context aggregation by dilated convolutions. Computer Science, In: CoRR, ArXiv, volume:abs/1511.07122.

[10] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L. (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. J. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 834-848.

[11] Chen, L.C., Papandreou, G., Schroff, F., Adam, H. (2017) Rethinking atrous convolution for semantic image segmentation. J. Computer Science, ArXiv, volume:abs/1706.05587.

[12] Fu J., Liu J., Tian H. (2019) Dual Attention Network for Scene Segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 3141-3149.

[13] Xie, S., Girshick, R., Dollár, P. Tu, Z. He, K., (2017) Aggregated Residual Transformations for Deep Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 5987-5995.

[14] Fergus, R. Taylor, G.W. Zeiler, M. D. (2011) Adaptive deconvolutional networks for mid and high level feature learning. In: International Conference on Computer Vision, Barcelona, Spain, pp. 2018-2025.

[15] Zhong, X., Tang J., Yepes A J. (2019) PubLayNet: Largest Dataset Ever for Document Layout Analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, pp. 1015-1022.

[16] Simonyan, K., Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. J. Computer Science, ArXiv, volume:abs/1409.1556.