J. Cent. South Univ. (2021) 28: 1765−1774
DOI: https://doi.org/10.1007/s11771-021-4731-9

Springer

End-to-end dilated convolution network for document image semantic segmentation

XU Can-hui(许灿辉)[1, 2], SHI Cao(史操)[1], CHEN Yi-nong(陈以农)[2]

1. School of Information Sciences and Technology, Qingdao University of Science and Technology, Qingdao 266061, China;
2. School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287-8809, USA

Abstract: Semantic segmentation is a crucial step for document understanding. In this paper, an NVIDIA Jetson Nano-based platform is applied for implementing semantic segmentation for teaching artificial intelligence concepts and programming. To extract semantic structures from document images, we present an end-to-end dilated convolution network architecture. Dilated convolutions have well-known advantages for extracting multi-scale context information without losing spatial resolution. Our model utilizes dilated convolutions with residual network to represent the image features and predicting pixel labels. The convolution part works as feature extractor to obtain multidimensional and hierarchical image features. The consecutive deconvolution is used for producing full resolution segmentation prediction. The probability of each pixel decides its predefined semantic class label. To understand segmentation granularity, we compare performances at three different levels. From fine grained class to coarse class levels, the proposed dilated convolution network architecture is evaluated on three document datasets. The experimental results have shown that both semantic data distribution imbalance and network depth are import factors that influence the document's semantic segmentation performances. The research is aimed at offering an education resource for teaching artificial intelligence concepts and techniques.

Key words: semantic segmentation; document images; deep learning; NVIDIA jetson nano

Cite this article as: XU Can-hui, SHI Cao, CHEN Yi-nong. End-to-end dilated convolution network for document image semantic segmentation [J]. Journal of Central South University, 2021, 28(6): 1765−1774. DOI: https://doi.org/10.1007/s11771-021-4731-9.

1 Introduction

For artificial intelligence (AI) education, various low-power internet of things (IoT) architectures, such as Galileo, Edison, and NVIDIA, have provided good platforms. They all have adequate capacity to perform complex computations [1, 2]. The GPU embedded SparkFun Jetbot is applied as our AI robotic platform, which is powered by NVIDIA Nano GPU board. These GPU embedded platforms have their advantages to use convolutional deep learning networks for various applications, such as face recognition, traffic sign recognition, pose estimation, semantic segmenation, and video processing. It is reported that NVIDIA

Foundation item: Project(61806107) supported by the National Natural Science Foundation of China; Project supported by the Shandong Key Laboratory of Wisdom Mine Information Technology, China; Project supported by the Opening Project of State Key Laboratory of Digital Publishing Technology, China

Received date: 2020-06-17; **Accepted date:** 2021-01-04

Corresponding author: SHI Cao, PhD, Lecturer; Tel: +86-532-88959809; E-mail: caoshi@yeah.net; ORCID: https://orcid.org/0000-0003-2748-5557

Jetson was used for low-power image recognition system with networks such as LeNet [3], Fast R-CNN, Yolo v3 and Mask-RCNN. Jetbot Nano is able to run networks in parallel. GPU acceleration made it possible for deep learning networks to work on these embedded platforms. To teach the robots read documents and build a document image application is one of the case studies for AI education.

Layout analysis and layout understanding, as two major parts of document processing, are separately dealt with in traditional manner. Layout analysis, also known as page segmentation, aims to segment the page geometrically into regions like text line, text block, etc. Layout understanding, often referred as logical understanding, takes the segmented regions as input for logical structure recognition of semantic labels like paragraph, section, figure, table, and list. These two parts are carried out on image documents or born-digital documents in a sequence, which makes the classification performance highly dependent on the front-end page segmentation results. Possible error in the previous segmenting stage tends to accumulate the misclassification into the second stage. To simultaneously segment and classify images at pixel level, semantic segmentation is often considered to be a better solution.

Recently, deep convolution neural networks have been proposed as a major solution for performing the tasks like image classification, object detection and image segmentation. Among various approaches, fully convolutional network (FCN) shows improvement on semantic segmentation task [4, 5]. It combines goals of page segmentation and semantic identification seamlessly, which is able to produce segmentation and labeling results in pixel level. There is attempt in introducing this pixel-level classification to document image processing area [6].

In this paper, we study semantic segmentation on document images. To implement AI on a Jetson Nano robot with limited processing capacity, we introduce image data preparation, training, optimizing and deployment for document semantic segmentation. Dilated convolution network architecture is proposed to represent the image features and predicting pixel labels. The influence of various granularities on semantic segmentation performance is discussed. To understand segmentation granularity, we compare performances in three different levels. From fine grained level to coarse class levels, the proposed dilated convolution network architecture is evaluated on three document datasets. The rest of the paper is organized as follows. Related studies on image document analysis are discussed in Section 2. The proposed end-to-end dilated convolution network architecture is established in Section 3. Experimental results are presented in Section 4. The conclusions are given in Section 5.

2 Related works

Traditionally, layout analysis was dedicated to differentiate text from non-text, followed by layout logic understanding to accomplish the task of recognizing logical classes like paragraph, figure, table, etc. For page segmentation, bottom-up and top-bottom methods are the major approaches in the existing studies [7−11]. Connected component (CC) based methods have dominated the document segmentation research field for decades. Low-level image features such as gradient, proximity, and texture are used to group text pixels or super pixels. CC is also popular among the leading segmentation algorithms in international conference on document analysis and recognition (ICDAR) segmentation competition [12]. As for logic understanding, various classification methods have been proposed to increase recognition accuracy. Support vector machine (SVM) is often used for multi-class classification in document labeling task [13−15]. LUONG et al [16] applied a linear chain based CRF model for document logical structure recovery. TAO et al [17] proposed a contextual model to label logical classes for digital-born PDF documents. With consideration of contextual information, both unary and binary features were utilized for classification [18, 19].

Connected component methods usually explore limited low-level features such as gradient and intensity. However, convolutional neural networks (CNN) is powerful in representing hierarchal features. Deep networks are able to naturally integrate low/mid/high-level features and classifiers. CNN has brought a series of improvement for document image segmentation and classification [6, 12]. Various CNN architectures, including VGGNet, ResNet, GoogLeNet, DeconvNet, etc, were often used in image document processing [20].

Among all these approaches, some researchers tend to use proposal strategies to segment document pages. CNN classifier is defined to classify proposals according to the deep convolutional features. Region CNN and its variations were applied in selective search. GIRSHICK [21] proposed a region proposal network (RPN), which was fast with the sharing convolution computation property. To adjust for document image processing, redesigning the region proposal methods was proven to achieve better performance [12, 22].

There is another group of studies aiming to predict pixel-wise class labels. Usually, CNN based image classification architecture was transferred to a fully convolutional network (FCN) for semantic segmentation [5]. Based on coarse feature map extracted from CNN, deconvolution was performed to obtain full resolution segmentation mask. To improve spatial accuracy, conditional random field (CRF) integrated with FCN could further improve the segmentation map by capturing fine local details [4, 23]. However, FCN has problems in labeling small size objects resulting in inconsistent labels due to fixed-sized receptive field. NOH et al [24] proposed a learning deconvolution network to solve this problem. Dilated residual network (DRN) network proposed by YU et al replaces subsampling layers by adding dilation, which could be applied for semantic segmentation [25].

Our model utilizes dilated convolutions with residual network to segment image document pages. Dilated convolutions have well-known advantages to extract multi-scale context information without losing spatial resolution. With exponential increase of receptive field, it can cover wide contextual field within the image thereby extracting better contextual features. According to human intuition, readers are able to visually segment the document page semantically with multi-scale contextual reasoning. Retaining the contextual information can disambiguate some mislabeled classes caused by using the local features. The consequent deconvolution is operated to produce full resolution prediction and pixel-wise class labels.

3 Proposed method

Intuitively, images can be visually segmented semantically with multi-scale contextual reasoning. Both local and contextual information are important to understand the document structure.

3.1 Dilated convolutional network

Dilation convolutional network is able to preserve spatial information [25]. Started with residual network (ResNet) architecture proposed by HE et al [26], our model is carried out with groups as well. For each group, a feature map is obtained after each layer. Let $I(x_1, x_2)$ denote the image, and $K(v_1, v_2)$ is the kernel, also known as the filter. The discrete convolution is denoted by $*$ operator. For each feature map, the convolution output is:

$$I(x_1,x_2)*K(v_1,v_2)=\sum_{v_1}\sum_{v_2} I(x_1-l_1v_1,x_2-l_2v_2)K(v_1,v_2) \tag{1}$$

The above formula represents two-dimensional dilation convolution. Here, l_1 and l_2 are called dilation factors. When $l_1=l_2=l$, a compact form of the output can be gained:

$$(I*K)\boldsymbol{x}=\sum_{\boldsymbol{v}} I(\boldsymbol{x}-l\boldsymbol{v})K\boldsymbol{v} \tag{2}$$

where $\boldsymbol{x}$ and $\boldsymbol{v}$ correspond to coordinates (x_1, x_2) and (v_1, v_2), respectively. When $l=1$, the formula has a familiar form of discrete convolution. Dilated convolution is also related with atrous convolution or convolution with holes [27]. By systematic dilation, receptive fields can be expanded exponentially without losing resolution. The dimension of the receptive field is a square of exponentially increasing size, with the form $(2^{i+2}-1)\times(2^{i+2}-1)$ for $i=1, \cdots, n-2$. $I_{i+1}=I_i*K$ and $l=2^i$. Given $i=0$, the receptive field is 3×3. Given $i=1$, $l=2$ and the receptive field is 7×7. Given $i=2$, $l=4$ and the receptive field is 15×15. Dilation of original ResNet can be reviewed as situation in dilated convolutional network with $l=1$. By increasing the receptive field at high layers, dilated convolutional network can remove the subsampling step which causes the reduction of spatial information.

As shown in Figure 1, the network architecture is arranged by groups. Each group maintains the

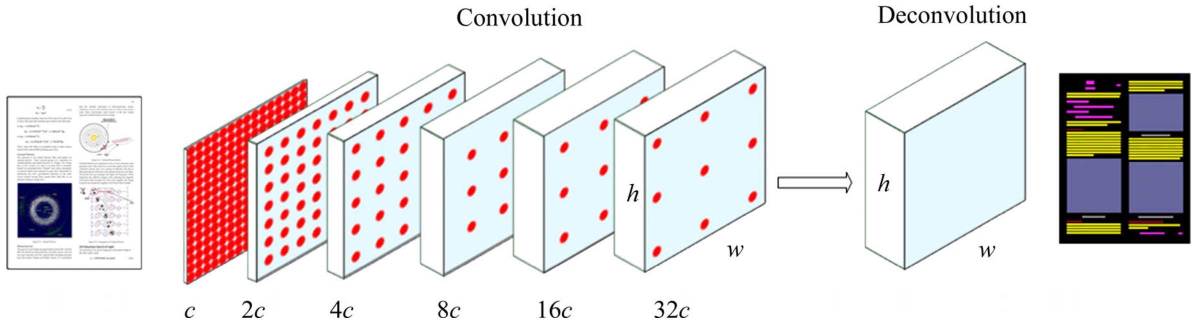


Figure 1 Architecture of dilated convolution network

same dilation operation as well as spatial resolution. Within each group, each rectangle represents a group of convolution, batch normalization, ReLU (Conv-BN-ReLU) operations. The dilation factor *l* equals 1 in early group layers, which is to keep the most information. And *c* is the channel dimension, which is also the number of feature maps in each group. Channel dimension is increased by the factor of 2. Residual connections are added and explored in different ways. *w* and *h* are the width and height of the feature map. The convolution part works as feature extractor to obtain multi-dimensional and hierarchical image features. And the consequent deconvolution is to produce full resolution segmentation prediction. The probability of each pixel decides its predefined class label.

Within convolution networks, successive pooling layers can obtain multi-scale contextual features with reduced resolution. To retain both multi-scale contextual reasoning and full resolution prediction, dilated convolution applies convolution network without pooling or subsampling to recover resolution. By replacing the pooling layers within the network, dilation convolution is applied to increase the output resolution. The consequent deconvolution is operated to produce full resolution prediction.

3.2 Network architecture

Dilated convolution is able to obtain long-ranged contextual information. Large dilation factor is useful for large object segmentation. For small objects, small dilation factor is more efficient. Designing dilated convolution with appropriate dilation factor is able to segmentation objects with various sizes [28]. Document image has its inherent features, especially structural layout with large area of block objects. Long-ranged information is of importance for page object detection and recognition. Our model takes colored document image as input. It has 8 groups. The early groups keep most information by applying regular convolution. Dilated convolution layers expanded the receptive fields to retain spatial information. Residual connections are added for first six groups. Convolutional filter kernel size is 3×3. The channel dimension is set to [16, 32, 64, 128, 256, 512, 512, 512]. And the dilation factors for DRN26 are set to [1, 1, 1, 1, 2, 4, 2, 1]. The receptive field of our model did not increase exponentially at higher layers. We applied ResNet network and removed the pooling layers which are replaced by dilation by a factor of 2. The original ResNet downsamples the input image by a factor of 32 in each dimension. Compared with ResNet, the dilation residual net (DRN) downsamples the input by a factor of 8. Our input images are scaled to 400×400. The output activation map of last group is 50×50, which retains more values than that of ResNet. For the network structure, two types of network have been evaluated on our datasets. We increased our network depth from 26 layers to 77 layers.

However, because the dilated kernel does not take continuous connected pixels, the output information obtained suffers checker-board like grid patterns. This effect, known as gridding artifact, is inappropriate for pixel-level dense prediction. It has been studied that removing residual connections at the top groups can eliminate gridding artifacts. Hence, dilation is added for the last two groups and their residual connections are removed.

3.3 Jetson Nano implementation

A Jetson Nano robot has 128 NVIDIA cores and costs under $300. It is affordable yet powerful robot for AI education. It is able to run various advanced networks. The Jetson Nano Developer Kit enables advanced neural networks to run in parallel for various applications in a low-power platform. A USB web camera is connected to the Nano board. Pytorch framework is used for deployment. HDMI interface is used for showing results.

4 Experimental results

4.1 Datasets for experiments

As is known that document ground truth datasets at pixel level are insufficient, there exist several possible solutions to this limitation when deep network training is concerned. YANG et al [6] made attempts in generating synthetic documents by scrapping data from internet and applying Latex. YI et al [12] used semi-automatic method to label ground truth. To explore the performance of our network architecture on document semantic segmentation, we investigate three datasets: Dataset I Marmot [17], Dataset II DSSE-200 [6], and Dataset III ICDAR2017 [29].

Dataset I Marmot, which selected from 35 English and Chinese books, has 244 image pages. It was also used in our previous work [17], which can be accessed through http://www.icst.pku.edu.cn/cpdp/sjzy. Our ground-truthing tool based on wxpython was able to label the document images at a given granularity. In this paper, we mark the document pages at fragment-level. Elaborative semantic logic labels are used to mark all the fragments. A set of 16 classes includes body text, title, figure, figure annotation, figure caption, figure caption continuation, list item, list item continuation, table cell, table caption, equation, page number, footer, header, note, and margins. All 244 images are divided into three groups for training, validation and testing with a ratio of 2:1:1. The data distribution among the 16 classes is highly imbalanced, which is also the case in other studies [12]. YANG et al [6] has designed a fewer classes within the document pages to improve the situation. In this dataset, body text class dominates over 50% of the fragments.

Dataset II DSSE-200 provides 201 labeled real document images for testing [6]. A large number of synthetic documents are used for training. The synthetic documents are generated from these manually labeled documents. This dataset has 6 classes, including figure, text, section, caption, list, and paragraph. The labeling in the ground truth contains some overlapping regions. All 201 images are divided into three groups for training, validation and testing with a ratio of 2:1:1. The data distribution among the 6 classes is also imbalanced. In this dataset, body text block class still dominates with 46% blocks.

Dataset III is from page object detection (POD) competition in conference ICDAR2017 [29]. It consists of 2417 document pages, which are selected from CiteSeer scientific papers. It includes 3 classes: figure, formula, and table. There are 9422 objects in total, with around 58% formulas, 31% figures and 11% tables.

4.2 Implementation details

Our network architecture is implemented in Pytorch. All the input images are scaled into 400 pixels in length. Increasing image size will cause greater computation. However, it results in more accuracy than smaller crop size. In other words, smaller crop size hurts the performance. The mean and standard deviation is calculated for preprocessing the data. Image-cropping and mirroring with horizontal flipping are used for training. No other data augmentation is added.

Popular pretrained models on natural images are not suitable for document images. We tried to use pretrained model of dilated residual network (DRN) on Cityscape dataset, and the results were not satisfactory. Hence, a convolutional network from scratch with random initialization is trained. High resolution label map images are used for end-to-end semantic segmentation. For feature map, each pixel is assigned a class label. Stochastic gradient descent (SGD) is applied for training, with mini-batch size of 16, learning rate of 10^{-3}, and momentum of 0.9. Batch normalization synchronization across multiple GPU is applied in our deep convolution network model for semantic segmentation. The network was trained for 200 iterations. The evaluation intersection-over-union (IoU) is used for evaluation. Similar to upsampling, up-convolutions are used to output full resolution prediction results.

As can be seen in Table 1, classes such as body, title, equation, figure annotation, figure caption, table cell, and page number have prediction results, among which body fragments have the highest average precision (AP) value. The rest of classes are not recalled due to their small number of fragments for training. The unbalance property among the classes deteriorates the segmentation performance.

Table 1 Semantic segmentation results on Database I Marmot

Class	Number of fragment	AP/%	
		DRN26	DRN77
Body	5460	55.31	55.78
Title	238	3.74	12.61
Equation	432	67.71	68.11
FigureAnnot	430	8.70	11.63
TableCell	2161	21.78	32.34
PageNum	226	17.30	11.49

The training process can be completed within 30 min. The loss can converge to lower than 0.3. The testing process can have the prediction result in 16 s for each image. No post processing is involved in the prediction process. Our model achieves segmentation as well as semantic prediction. Low IoU between candidate and ground truth is also the reason for the low mAP results. Segmenting granularity to fine level of 16 classes is not appropriate for semantic segmentation.

In Table 2, the segment granularity is set to 6 classes. Except caption class, all other classes have prediction results, among which the figure block has the highest average precision (AP) value. Compared to blocks with large area, caption blocks are generally smaller in area, and their training samples are limited. Within 30 min, we can complete training process. No post processing is involved in the prediction process. The interesting fact is that the figure class can be predicted with much better results when the text line fragments are aggregated to text blocks. By reducing the semantic classes, the granularity is coarser and the unbalance problem is alleviated to a certain point.

Table 3 summarizes the segmentation results of three semantic classes: figure, formula, and table. Unbalance is no longer that severe. This task is more akin to extracting objects from the entire document pages. All three classes have prediction results, among which table objects have the highest AP value. Training process takes about 3 h. No extra post processing is involved in the prediction process, without rounding pixels to bounding boxes. Compared to the results from POD competition ICDAR2017 [29], our model can achieve prediction result of 71.97% without further post processing techniques or extra auxiliary features.

Table 2 Semantic segmentation results on Database II DSSE-200

Class	Number of block	AP/%	
		DRN26	DRN77
Text	1188	13.46	16.42
Figure	288	41.96	41.16
Section	658	0.0	4.89
Caption	126	0.0	0.0
List	217	0.0	0.002
Table	79	26.11	32.01

Table 3 Semantic segmentation results on Database III ICDAR2017

Class	Number of object	AP/%	
		DRN26	DRN77
Figure	2955	64.48	76.24
Formula	5447	45.49	60.74
Table	1020	56.82	78.94
This study	9422	55.60	71.97

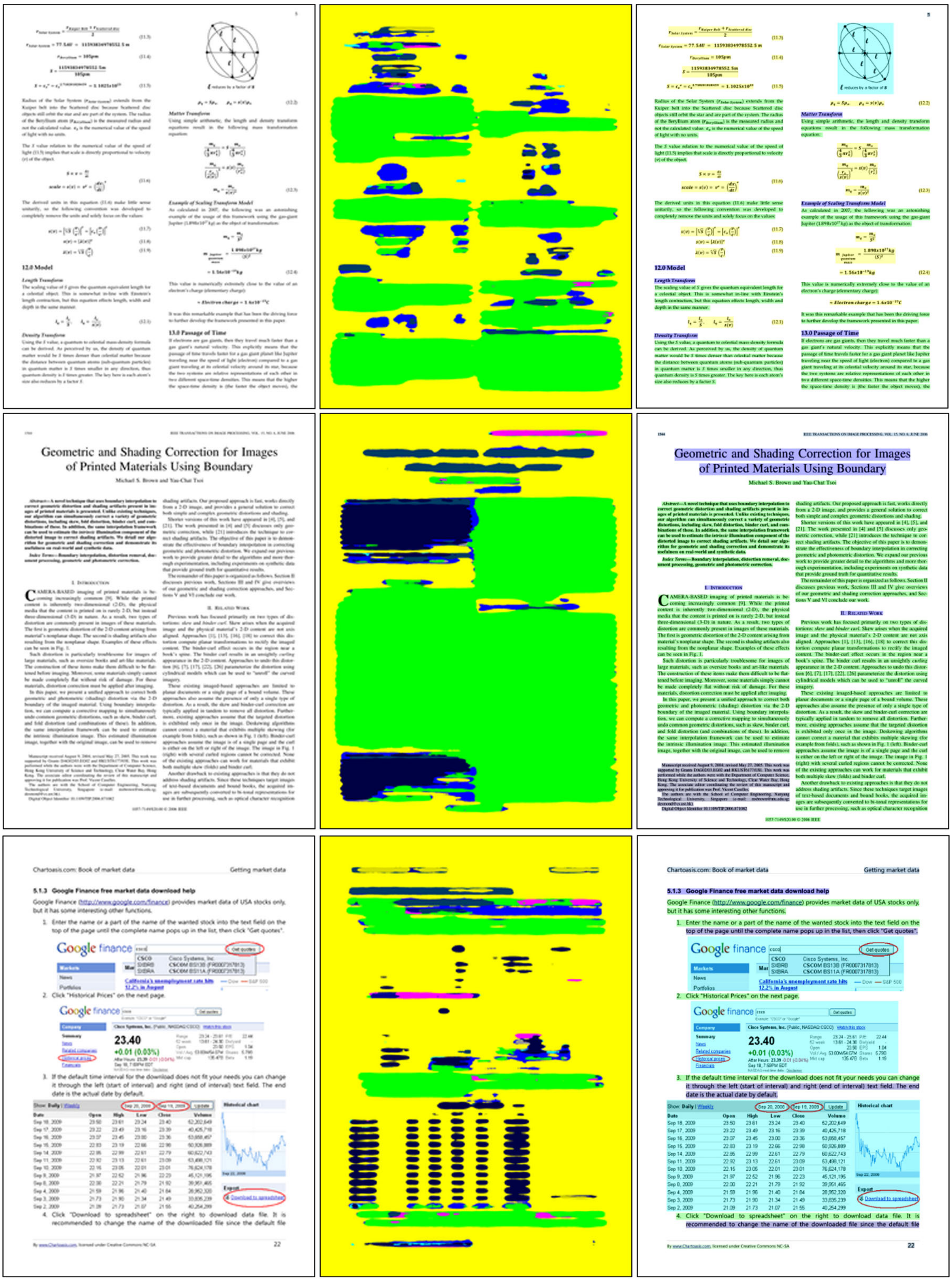
Table 4 shows the comparison results of various methods, including bounding box based detection methods and pixel-wise semantic segmentation methods. Both Fast RCNN [21] and Faster RCNN [30] output bounding boxes for document classes. These two methods were originally implemented for natural image object detection. Yi SPP based method [12] adapted the region proposal network to document object detection, and it performs better on small objects, such as formulas. SSD is not good at detecting small objects, being unable to detect any formula [31]. Our dilated convolution network based method has general good feature representation for

Table 4 Comparison of different methods on segmentation results of Database III ICDAR2017

Method	AP/%		
	Figure	Formula	Table
Fast RCNN [21]	46.3	14.5	58.2
Faster RCNN [30]	50.3	10.8	61.1
SSD [31]	34.7	0	43.5
HAO et al [32]	—	—	70.1
LIN et al [13]	68.6	—	—
YI et al [12]	66.8	79.7	84.2
This study	76.24	60.74	78.94

page document images. And it achieves AP of 76.24% on figure class in pixel-wise level, without any further post processing.

To understand segmentation granularity, we compared performances at different levels. When the granularity is in the range of fine grained classes, Dataset I Marmot logically divided page objects into 16 classes. Intuitively observing the segmentation results, Figure 2 shows three sampled document images along with the segmentation results from



(a) (b) (c)

Figure 2 Example documents and their semantic segmentation results: (a) Document pages from Database I Marmot; (b) Segmenting results of the document pages; (c) Segmentation groundtruth

Dataset I Marmot. The mean average precision (mAP) on Dataset I Marmot is relatively low, and the segmented text fragment results are visually acceptable. Being the major class, text fragments are well trained and generally easy to be detected. However, small objects or objects with small amount of training data disappear among the detection results. For coarse class level, as is shown in Figure 3, the visual segmenting results are surprisingly good, and the data imbalance problem is alleviated. When the segmentation granularity is in block level, AP performance on page object detection (POD) competition ICDAR2017 boosts for each class.

The granularity is an important factor to consider in document semantic segmentation. Another fact is that when the depth of network goes deeper, the model can achieve better results. Unbalances problem is obvious in document pages. The majority class is body text in Dataset I and Dataset II. The results have shown that the unbalance property heavily reduces the prediction accuracy. Many categories have no prediction results due to insufficient training samples and severe bias of data class distribution.

5 Conclusions

In this study, a dilated convolution network was proposed to semantically segment both images and documents. The experiments on the three datasets were conducted and the results revealed certain insights of segmentation granularity and network depth, including:

1) Concerning segmentation granularity, intensive classes with highly imbalanced data distribution would deteriorate the prediction performance.

2) When the small fragments were aggregated to relatively large blocks, the segmentation granularity got coarse, and data imbalance problem was alleviated. Hence, the performance became better.

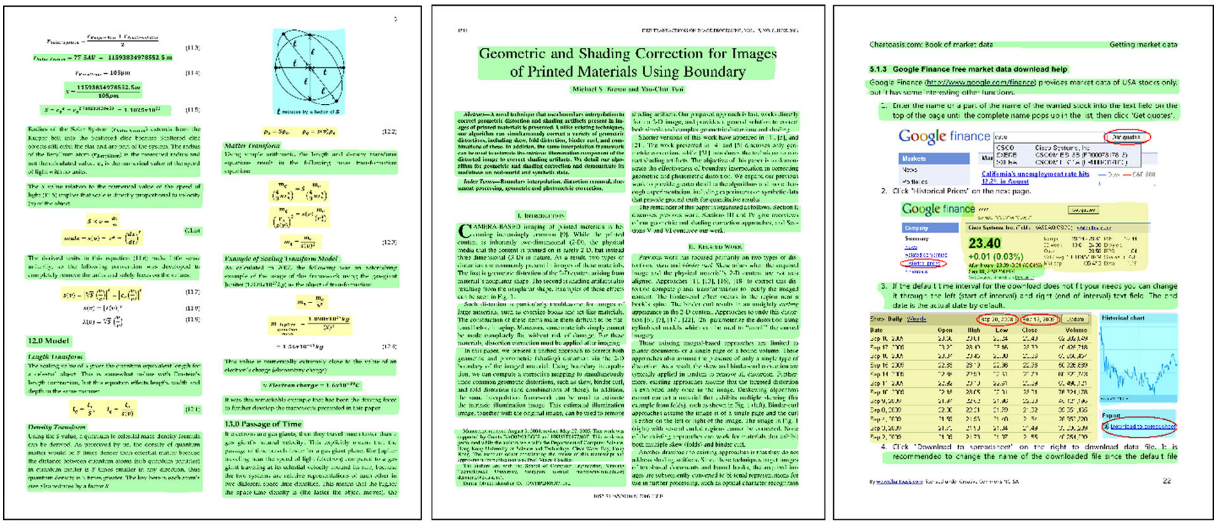
3) With the increasing network depth, the model could represent the features better and hence achieve better results. Deep learning feature representation of page object at hierarchical level will be one of our future research areas.

Contributors

XU Can-hui proposed the concept, implemented the experiments and wrote the first draft of manuscript. SHI Cao conducted the literature review, collected the document ground truth datasets and edited the draft of the manuscript. CHEN Yi-nong edited the draft of manuscript.

Conflict of interest

XU Can-hui, SHI Cao and CHEN Yi-nong declare that they have no conflict of interest.



(a) (b) (c)

Figure 3 Semantic segmentation results from Database I Marmot with aggregated page objects of four classes, including text block, figure, table and math (Text fragments are colored in green. Maths are in yellow. Figures are in cyan. Tables are in middle washed yellow)

References

[1] CHEN Yi-nong, ZHOU Zhi-zheng. Service-oriented computing and software integration in computing curriculum [C]// 2014 IEEE International Parallel & Distributed Processing Symposium Workshops. Phoenix, AZ, USA: IEEE. 2014: 14792480. DOI: 10.1109/IPDPSW.2014.127.

[2] CHEN Yi-nong, DE LUCA G. VIPLE: Visual IoT/robotics programming language environment for computer science education [C]// 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Chicago, IL, USA. IEEE, 2016: 963−971. DOI: 10.1109/IPDPSW.2016.55.

[3] HAN Yan, ORUKLU E. Traffic sign recognition based on the NVIDIA Jetson TX1 embedded system using convolutional neural networks [C]// 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). Boston, MA, USA: IEEE, 2017: 184−187. DOI: 10.1109/MWSCAS.2017.8052891.

[4] CHEN L C, PAPANDREOU G, KOKKINOS I, MURPHY K, YUILLE A L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834−848. DOI: 10.1109/TPAMI.2017.2699184.

[5] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640−651. DOI: 10.1109/TPAMI.2016.2572683.

[6] YANG Xiao, YUMER E, ASENTE P, KRALEY M, KIFER D, GILES C L. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 4342−4351. DOI: 10.1109/CVPR.2017.462.

[7] DRIVAS D, AMIN A. Page segmentation and classification utilising a bottom-up approach [J]. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, 2: 610−614. DOI: 10.1109/ICDAR.1995.601970.

[8] SIMON A, PRET J C, JOHNSON A P. A fast algorithm for bottom-up document layout analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(3): 273−277. DOI: 10.1109/34.584106.

[9] HA J, HARALICK R M, PHILLIPS I T. Recursive X-Y cut using bounding boxes of connected components [C]// Proceedings of 3rd International Conference on Document Analysis and Recognition. Montreal, QC, Canada: IEEE, 1995: 952−955. DOI: 10.1109/ICDAR.1995.602059.

[10] CAI Deng, YU Shi-peng, WEN Ji-Rong, MA Wei-ying. Vips: A vision-based page segmentation algorithm [EB/OL]. [2003-11-01]. https://www.microsoft.com/en-us/research/publication/vips-a-vision-based-page-segmentation-algorithm/.

[11] KISE K, SATO A, IWATA M. Segmentation of page images using the area voronoi diagram [J]. Computer Vision and Image Understanding, 1998, 70(3): 370−382. DOI: 10.1006/cviu.1998.0684.

[12] YI Xiao-han, GAO Liang-cai, LIAO Yuan, ZHANG Xiao-de, LIU Run-tao, JIANG Zhuo-ren. CNN based page object detection in document images [C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan: IEEE, 2017: 230−235. DOI: 10.1109/ICDAR.2017.46.

[13] LIN Xiao-yan, GAO Liang-cai, TANG Zhi, BAKER J, SORGE V. Mathematical formula identification and performance evaluation in PDF documents [J]. International Journal on Document Analysis and Recognition (IJDAR), 2014, 17(3): 239−255. DOI: 10.1007/s10032-013-0216-1.

[14] XU Can-hui, TANG Zhi, TAO Xin, LI Yun, SHI Cao. Graph-based layout analysis for PDF documents [C]//Proc SPIE 8664, Imaging and Printing in a Web 2 0 World IV, 2013, 8664: 866407. DOI: 10.1117/12.2005608.

[15] XU Can-hui, TANG Zhi, TAO Xin, SHI Cao. Graphic composite segmentation for PDF documents with complex layouts [C]// Document Recognition and Retrieval XX, 2013, 8658: 86580E. DOI: 10.1117/12.2003705.

[16] LUONG M T, NGUYEN T D, KAN M Y. Logical structure recovery in scholarly articles with rich document features [M]// Multimedia Storage and Retrieval Innovations for Digital Library Systems. IGI Global, 2012: 270−292. DOI: 10.4018/978-1-4666-0900-6.ch014.

[17] TAO X, TANG Z, XU C. Contextual modeling for logical labeling of PDF documents [J]. Computers & Electrical Engineering, 2014, 40(4): 1363−1375. DOI: 10.1016/j.compeleceng.2014.01.005.

[18] TAO Xin, TANG Zhi, XU Can-hui, WANG Yong-tao. Logical labeling of fixed layout PDF documents using multiple contexts [C]// 2014 11th IAPR International Workshop on Document Analysis Systems. Tours, France: IEEE, 2014: 360−364. DOI: 10.1109/DAS.2014.54.

[19] DELAYE A, LIU Cheng-lin. Contextual text/non-text stroke classification in online handwritten notes with conditional random fields [J]. Pattern Recognition, 2014, 47(3): 959−968. DOI: 10.1016/j.patcog.2013.04.017.

[20] VINCENT N, OGIER J M. Shall deep learning be the mandatory future of document analysis problems? [J]. Pattern Recognition, 2019, 86: 281−289. DOI: 10.1016/j.patcog.2018.09.010.

[21] GIRSHICK R. Fast R-CNN [C]// 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1440−1448. DOI: 10.1109/ICCV.2015.169.

[22] TIAN Zhi, HUANG Wei-lin, HE Tong, HE Pan, QIAO Yu. Detecting text in natural image with connectionist text proposal network [M]// Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 56−72. DOI: 10.1007/978-3-319-46484-8_4.

[23] CHEN L C, PAPANDREOU G, KOKKINOS I, MURPHY K, YUILLE A L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834−848. DOI: 10.1109/TPAMI.2017.2699184.

[24] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation [C]// 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1520−1528. DOI: 10.1109/ICCV.2015.178.

[25] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 636−644. DOI: 10.1109/CVPR.2017.75.

[26] HE Kai-ming, ZHANG Xiang-yu, REN Shao-qing, SUN Jian. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770−778. DOI: 10.1109/CVPR.2016.90.

[27] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions [EB/OL].[2015-11-23]. https://arxiv.org/abs/1511.07122.

[28] CHEN L C, PAPANDREOU G, SCHROFF F, ADAM H. Rethinking atrous convolution for semantic image segmentation [EB/OL].[2017-06-17]. https://arxiv.org/abs/1706.05587.

[29] GAO Liang-cai, YI Xiao-han, JIANG Zhuo-ren, HAO Lei-peng, TANG Zhi. ICDAR2017 competition on page object detection [C]// 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan: IEEE, 2017: 1417−1422. DOI: 10.1109/ICDAR.2017.231.

[30] REN Shao-qing, HE Kai-ming, GIRSHICK R, SUN Jian. Faster R-CNN: Towards real-time object detection with region proposal networks [C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE, 2016: 1137−1149. DOI: 10.1109/TPAMI.2016.2577031.

[31] LIU W, ANGUELOV D, ERHAN D, SZEGEDY C, REED S, FU C Y, BERG A C. SSD: Single shot multibox detector [C]// European Conference on Computer Vision, Amsterdam: Springer, 2016: 21−37. DOI: 10.1007/978-3-319-46448-0_2.

[32] HAO Lei-peng, GAO Liang-cai, YI Xiao-han, TANG Zhi. A table detection method for PDF documents based on convolutional neural networks [C]// 2016 12th IAPR Workshop on Document Analysis Systems (DAS). Santorini, Greece. IEEE, 2016: 287−292. DOI: 10.1109/DAS.2016.23.

(Edited by ZHENG Yu-tong)

中文导读

基于膨胀卷积网络的端到端文档语义分割

摘要：本文采用膨胀卷积网络，实现端到端从文档图像中提取语义结构。膨胀卷积的优势在于提取多尺度上下文信息的同时，并不会损失空间分辨率。该模型使用带残差的膨胀卷积网络提取图像特征，并预测每个像素的类别标签。卷积部分作为特征提取器，能够获得多维度层级图像特征，反卷积部分输出全分辨率的语义预测结果。每个像素的概率值决定其语义类别标签。为了更好地理解分割粒度级别，实验设计了 3 组不同分割粒级数据集的测试。从文档细粒度到粗粒度级别的分割实验结果表明，语义数据分布的不平衡特点和网络深度都是影响该网络模型的重要因素。该模型可测试于人工智能教育平台英伟达 Jetson Nano 机器。

关键词：语义分割；文档图像；深度学习；英伟达 Jetson Nano