Multi-scale Gated Fully Convolutional DenseNets for semantic labeling of historical newspaper images

Yann Soullard[a,*], Pierrick Tranouez[b], Clément Chatelain[b], Stéphane Nicolas[b], Thierry Paquet[b]

[a] *LETG-Rennes, University of Rennes, 35000 Rennes, France*
[b] *LITIS lab, Normandie University, Universite Rouen Normandie, INSA Rouen Normandie, 76800 Saint Etienne du Rouvray, France*

Abstract

Historical newspaper image analysis is a challenging task due to the complex layout of newspapers and its variability among collections. While traditional approaches are rule-based methods with many successive steps, recent works show that deep learning approaches can be successfully used to provide a pixel labeling of the various fields occurring in a page. This allows the automatic extraction of the document structure and accessing the different semantic entities. Recent improvements proposed to strengthen convolutional neural network capacities such as gated mechanism may also apply well to to task at end. In this respect, we propose a fully convolutional neural network architecture (FCN) that outputs a pixel-labeling of the various semantic entities that occur in historical newspaper images. Our model is based on a novel Multi-Scale Gated Block architecture (MSGB), made of dense connections and gating mechanisms that handle a multi-scale analysis of the input image with self-attention. Evaluations conducted on 4 historical newspaper datasets including up to 11 semantic classes show that our proposition outperforms standard FCN architectures.

Keywords: Historical newspapers, Fully Convolutional Networks, Multi-scale Gating Bock, Densely connected network, Gating mechanism, pixel labeling

1. Introduction

1.1. Aims

Historical newspapers are periodical publications containing many news, articles and events from a variety of fields. Information retrieval in historical newspapers can be useful in many fields of research such as history, geography, social sciences or genealogy.

Looking at the past competitions carried out on newspaper datasets, page segmentation and region identification is considered focusing on few physical classes such as, texts, graphics, separators and images (Antonacopoulos et al. (2013)). However, in order to understand the whole structure of a newspaper page, and extract its semantic components, logical or semantic labels such as captions, headlines and sub-headlines, or advertisement are needed, as exemplified by the datasets provided by Gallica from the Europeana Newspapers project[1] or by the Luxembourg National Library[2]. Introducing such semantic labels in additions to the physical labels increases the complexity of the task in two respects: 1- as the number of classes increases, the difficulty gets higher; 2- semantic labels may sometime require a much larger context to be desambiguited than physical labels.

For instance, in Figure 1, there are headlines or titles of sections or articles appearing at the same hierarchi-

*Corresponding author:
Email address: yann.soullard@univ-rennes2.fr (Yann Soullard)

[1] api.bnf.fr/documents-de-presse-numerises-en-mode-article-du-projet-europeana-newspapers
[2] data.bnl.lu/data/historical-newspapers/

cal level in the newspaper but that are printed with quite different typographical/physical features. It thus makes title level prediction impossible if solely based on local/physical features. The same goes for large heterogeneous areas like advertisements, which can be composed of text, graphics and images. Furthermore, compared to natural scene images, newspapers are composed of many local details such as graphical horizontal or vertical separators, punctuation marks, small characters, but also much larger patterns such as columns, tables or images. Last but not least, some historical newspapers have been printed on very large sheets of paper leading to very large images when scanning at the required resolution of at least 300 dpi to capture every details. Therefore, in order to capture and identify local details as well as large informative patterns, a multi-scale analysis is required.

1.2. Proposal

State of the art approaches for document structure analysis (Tranouez et al. (2015); Coüasnon and Lemaitre (2017)) combine machine learning and rule-based methods. Such methods can perform well using only few examples to tune the system. However, they have some drawbacks: i) they introduce many processing steps, such that an error at one step can impact the system performance in subsequent stages; ii) they generally have many hyper-parameters which make the system setting data-dependent; iii) they can fail in special cases on images with strong degradation.

Recently, deep learning methods have been proposed with success for semantic segmentation of natural scene images (Long et al. (2015)) or text-line detection in document images (Diem et al. (2017); Renton et al. (2018)). Fully Convolutional Networks (FCN) have been successfully introduced in order to label pixels on historical documents images (Xu et al. (2018); Oliveira et al. (2018)). Such approaches predict pixel's semantic classes in one step. Very recently, a number of advances have been proposed to even strengthen convolutional network capacities. For instance, densely connected architectures (Huang et al. (2017)) improve feature propagation inside the network while gating mechanisms (Yousef et al. (2018); Ingle et al. (2019)), so called self-attention mechanisms, select relevant features at the output of a layer. Such improved operators allow improving the system's capacities while maintaining almost a constant number of parameters in the architecture.

Figure 1: Difficulties/variability in newspaper semantic pixel labeling (from *Le Matin* newspaper, 08/17/1937 page 3). Left: numerous titles where some same logical level appears with different sizes and fonts. Right: an advertisement looking like an article.

Inspired by these recent advances in deep learning, this paper presents a fully convolutional neural network architecture based on a multi-scale gating mechanism combined with dilated convolutions and dense connections to achieve pixel labeling in the relevant context. The goal is to face the challenging problem of multi-scale analysis by combining features at different scales to make a decision. In this way, we propose a self-attention mechanism thanks to a novel multi-scale gate block which selects the relevant features and scales to predict the semantic class of each pixel. We evaluate our architecture on historical newspapers images in order to detect regions of interest

that appear at different scales, and pertain to up to 11 different semantic classes e.g. advertisement, images, different levels of titles and text blocks.

The rest of the paper is organized as follows: Section 2 introduces the related work on newspaper analysis and historical document layout analysis; In section 3, we present the fully convolutional network that we propose, with a multi-scale gate block; Then, we present and discuss experimental results using several network architectures applied on historical newspaper's images datasets (section 4) before concluding.

2. Related work

In this section we present related works dedicated to historical newspaper analysis. As there is only a few works specifically dedicated to historical newspapers analysis, we first present, more generally, an overview of methods dedicated to historical document layout analysis.

2.1. Historical Document image analysis

Generally, historical document image analysis has to cope with the many possible degradations that may occur due to aging of ink and paper. Moreover historical documents generally exhibit more irregular layouts than modern documents, up to being purely handwritten. These many sources of variability make historical document image analysis much more difficult than contemporary documents'. This is why machine learning approaches are very promising as they can cope with the variability of the input data, if sufficient labelled data can be provided to train the system.

When no ground truth is available, rule-based systems are the default approach. For instance, Lemaitre et al. (2018) presented a rule-based system based on keyword detection to extract handwritten fields in old pre-printed registers. Additionally, Bukhari et al. (2018) made use of the percentile based binarization method and multi-resolution morphology operations to segment medieval documents.

These last few years, deep learning approaches have been used with success for historical document layout analysis. Even if they generally require large amount of training examples, deep learning approaches can handle several tasks at the same time, ranging from physical to logical structure analysis. Quirós (2018) proposed a two-stage method for both physical and logical analysis. First, a pixel-level classification is performed using an Artificial Neural Network. Secondly, a zone segmentation and baseline detection are done using a contour extraction algorithm. Another work proposed the use of a Convolutional Neural Network (CNN) applied on superpixels (Chen et al. (2017)). In addition, Moysset et al. (2018) proposed a convolutional network with 2D-LSTM layers allowing to have more context than from the receptive fields of the CNN only. The network predicts area of interest such as boxes, corners or left sides.

Fully Convolutional Networks (FCN) have also been used with success for such a task by Wick and Puppe (2018), Xu et al. (2018) and Oliveira et al. (2018). FCN have many advantages compared to other deep learning approaches as they have generally less parameters (as there is no fully-connected layer) and they produce a pixel-level labeling. For instance, Oliveira et al. (2018) proposed a generic approach for document segmentation. The system is a FCN based on a ResNet architecture, providing good results on many tasks such as baseline detection, photo extraction or ornament detection. Xu et al. (2018) proposed a multi-task layout analysis using a FCN, where both page segmentation, text line segmentation and baseline detection are performed simultaneously.

2.2. Newspaper image analysis

Antonacopoulos et al. (2013) presented a comparative study of historical newspaper layout analysis, applied on newspapers coming from the IMPACT dataset. The proposed approaches generally deal with binarized images. Several pre-processing techniques are applied, such as skew correction and border deletion. Then, a bottom-up approach is generally used to aggregate connected components and to detect textual regions and graphical separators. More recently, Vasilopoulos and Kavallieratou (2017) proposed a method based on contour classification and morphological operations to find separators and extract text regions and titles.

The method proposed by Sfikas et al. (2016) is made up of two steps: the first one detects text, title regions and graphical separators. The second step merges some regions to get articles. The number of text regions and title regions are respectively estimated using a Bayesian Gaussian Mixture Models based clustering technique. In Riedl

et al. (2019), the authors presented a clustering-based article identification using word embeddings. While making use of a textual embedding to find articles seems promising, the system may be strongly affected by the OCR quality.

Besides, Tranouez et al. (2015) designed a software package called PIVAJ, dedicated to historical newspaper analysis. First, a layout analysis is performed through pixel-level classification using Conditional Random Fields and Random forests. Then, a grid providing a reading order, is created thanks to the extracted separators. Text blocs are then labeled and used to extract sections and articles spreading on multiple pages. This software package has been applied with success on digitized Finnish historical newspapers (Kettunen et al. (2019)).

3. Fully convolutional network with multi-scale gate block

Our fully convolutional model is based on a novel multi-scale gating block (MSGB) combining the benefit of the multi-scale analysis and a gating mechanism to provide a local and a global analysis of the image by self-attention. The proposed architecture also benefits from dilated convolutions and dense connections. We discuss these specific characteristics in the next sections.

3.1. Overall Architecture

The overall architecture is described in Figure 2 (a) and a convolutional block is detailed in Figure 2 (b). The multi-scale gate block is detailed later (section 3.3 and Figure 3).

The overall architecture is a Fully Convolutional Network (FCN). FCN are convolutional networks without dense layer. Removing this dense layer has many advantages: i) the number of parameters is highly reduced; ii) FCN can deal with variable input size; iii) the spatial information is kept inside the network: we can produce a pixel labeling.

Most FCN architectures are based on traditional convolutional operation mixing convolutional and pooling layers which reduces the input resolution. In such a case, the input resolution must be reconstructed to get pixel labeling. This is generally done using transposed convolution or up-scaling as in Long et al. (2015) and Oliveira et al. (2018). In this paper, we propose to use dilated convolutions which have been used with success for text-line identification (Renton et al. (2018)) and semantic segmentation (Yu and Koltun (2015)). Dilated convolutions can be seen as a generalization of the standard convolution. Using dilated convolutions has many advantages compared to traditional architectures: the input resolution is not decreased (avoiding the reconstruction step) and the features are never summarized (as with pooling layers). Here, the input resolution is kept throughout the network as we use a stride of 1 and padding to solve the border effects.

Our convolutional block begins by a 1×1 convolutional layer which acts as a feature pooling. Then Batch Normalization is used before a ReLU activation to increase convergence speed. Spatial dropout (Tompson et al. (2015)) is used to drop entire feature maps. As stated in Tompson et al. (2015), spatial dropout is preferred to standard dropout as FCN tends to exhibit strong spatial correlations and standard dropout fails in this setting. Then, two similar convolutional part (conv + BN + ReLU) are used with 3×3 kernels.

Several convolutional blocks are stacked in a dense architecture setting. Similarly to traditional FCN architecture and as in Renton et al. (2018), the dilatation rate is first increased between two blocks (as in an encoding part) and then decreased (as in the reconstruction part). Then the output of each block is given to a multi-scale gate block that provides the network output. To the next, we will discuss more in details dense nets and our proposed multi-scale gate block.

3.2. DenseNet

In densely connected networks proposed by Huang et al. (2017) (DenseNet) the output of each layer is given as input of the following layers. Such an architecture has many advantages: i) it reduces the vanishing-gradient problem, as there is more direct connections with the network output; ii) features can be reused at any step in the network; iii) the number of parameters is highly reduced compared to traditional Residual Networks (He et al. (2016)).

In a standard architecture, dense connections are used in a convolutional block due to pooling layers that reduces the input resolution and prevents feature propagation. This is a major drawback since dense connections

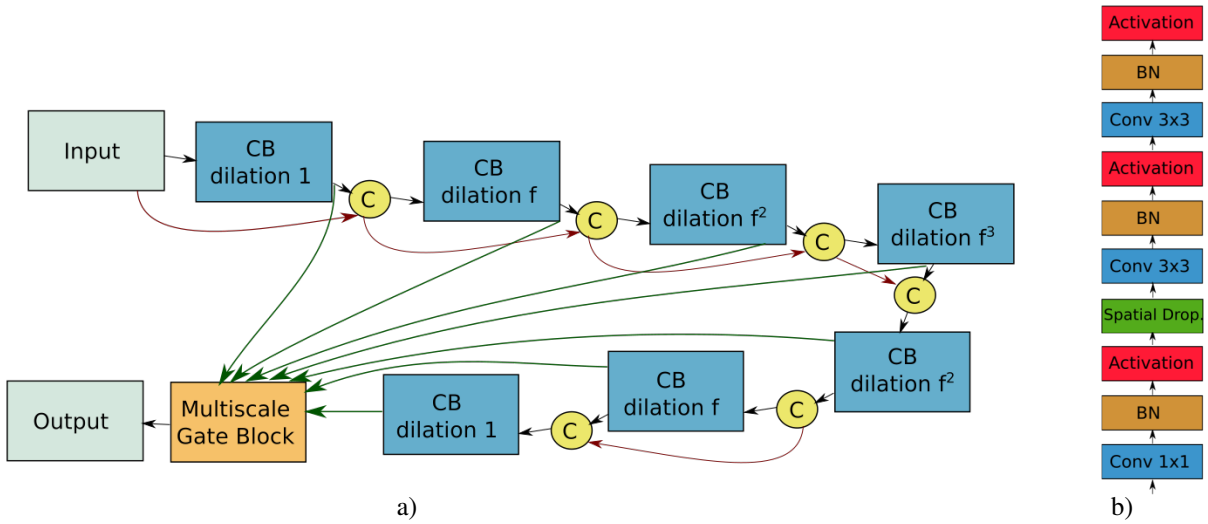

Figure 2: **a):** Overall network architecture **b):** Convolutional block. The overall network architecture is composed of Convolutional Block (CB) with a dilation factor f and C is a concatenation. A convolutional block is made of convolutional layer with a $k \times k$ kernel (k=1 or 3) providing 64 feature maps, Batch Normalization (BN), ReLU activations and spatial dropout.

can be far from the network output, therefore limiting the positive impact on the vanishing gradient problem. Here, due to the use of dilated convolutions, dense connections can be applied between blocks (red connections on Figure 2), allowing deeper connections and a better feature propagation among the different scales of the analysis.

3.3. Multi-scale gate block

Our multi-scale gate block is illustrated in Figure 3. The goal of this block is twofold. First, to consider features from various scales to improve the decision step: it should be highly beneficial with newspaper images as there are regions of interest that appear with different scales. Second, to focus the decision on relevant features. This can be seen as a self-attention mechanism inside the network. In addition, this block should help the model to recognize infrequent classes by combining a local and contextual analysis with a gating mechanism.

Our gating mechanism is inspired by recent works as Yousef et al. (2018) and Ingle et al. (2019). We propose to split the feature maps in two parts. The first one refers to a traditional *tanh* activation over the input feature maps at a particular scale, while the second one refers to the gate that acts as a self-attention mechanism over the first part. The gating operation is modeled using a *sigmoid* activation. Both parts are then multiplied to get the output of the gate block.

Let x be the input of the convolutional layer inside the gate block, x is of dimension $H \times W \times C^I$ where H, W and C^I relate to the height, width and number of channels respectively. One defines C^O the number of output channels (which is here the same as in input, i.e. $C^I = C^O$) and w denotes the filters which are here of dimension $1 \times 1 \times C^I \times C^O$ (as there is a 1×1 kernel). The input size is preserved on the output but as we apply a 1×1 convolution, there is no need for padding. Let i, j and k be the index in the output feature map, i.e. $1 \leqslant i \leqslant H$ and $1 \leqslant j \leqslant W$ and $1 \leqslant k \leqslant C^O$. Equation 1 gives the gate block operation:

$$y[i,j,k] = \underbrace{\tanh(z[i,j,k])}_{\textit{traditional output}} \times \underbrace{\sigma(g[i,j,k])}_{\textit{the gate}} \tag{1}$$

$$\text{where } z[i,j,k] = \sum_{c=1}^{C^I} w[c,k]x[i,j,c] + b[i,j,k]$$

$$\text{and } g[i,j,k] = \sum_{c=1}^{C^I} w[c,C^O+k]x[i,j,c] + b[i,j,C^O+k]$$

where b is the bias, z and g relate to the traditional and gate part respectively, tanh the hyperbolic tangent func-

tion and σ the sigmoid function. Finally, the outputs of every gate block are concatenated and given as the input of a final convolution layer followed by a softmax activation. In prediction, each pixel is assigned the class with the highest probability in output of the softmax activation.

3.4. Model training

The network architecture is composed of 7 convolutional blocks for a total of 21 convolutional layers before the multi-scale gate block. We use a dilation factor of 2, 3x3 convolution kernels and dropout of 0.2 (i.e. 20% of feature maps are dropped during training). Every convolutional layer provides 64 feature maps. In the multi-scale gate block, there are thus 128 channels that are split in two parts. The number of output channels is the number of classes and it depends on the task. The network is trained using the categorical cross-entropy loss function on batch size of 1 and we use RMSprop for gradient optimization with a learning rate of 10^{-4}. The number of epochs is equal to 60 and we report the results achieved on the test set, by selecting the best parameters on the validation set.

4. Experiments

This section presents the evaluated network architectures (section 4.1), the datasets (section 4.2) and the experimental results (section 4.3).

4.1. Network architectures

The proposed architecture based on MSGB described in Figure 2 and in 3.4 is compared with two other models:

- The first architecture (Dilated FCN called FCN) is the reference architecture: it is inspired by the FCN model based on dilated convolutions defined by Renton et al. (2018) which has been used with success for text-line identification. For a fair comparison with the proposed architecture based on MSGB, it is made of 7 convolutional blocks and the number of feature maps in each convolutional block is fixed to 64 and never increased.

- The second architecture (Dense FCN called DFCN) has a similar architecture than the proposed network without the multi-scale gate block (see Figure 2).

Table 1: Number of trainable parameters for every FCN architectures.

	Dilated FCN	Dense FCN	MSGB FCN
Parameters	545k	610k	668k

Table 2: Historical newspapers datasets: Le Matin (LM), GBNLA (GB), Luxemburger Wort (LW) and L'Indpendance Luxembourgeoise (IL).

	LM	GB	LW	IL
# newspapers	1	6	1	1
period	1885 to 1937	–	1877	1877
# classes	6	4	12	12
# in training	801	66	76	980
# in validation	100	12	12	120
# in test	100	24	16	120

Dilated FCN and Dense FCN architectures allow to observe the benefit of using a dense architecture and the multi-scale gated block. Table 1 shows the number of parameters per model. Let us emphasize that the proposed models are rather deep, but are reasonably light thanks to i) the fully convolutional nature of the proposed architecture that avoids heavy dense layers, ii) the denseNet that allows to reduce the number of feature maps, and iii) the use of 1×1 convolutions that also limits the number of channels. For comparison, the MSGB architecture is 200 times as small as VGG16 from Simonyan and Zisserman (2014) and a dilated FCN architecture as in Renton et al. (2018), where the number of channels would have been increased by 2 between 2 blocks, would give more than 10 millions of parameters.

4.2. Historical newspaper datasets

We experiment these architectures on 4 historical newspapers datasets: Le Matin, GBNLA and two from the Luxembourg (Table 2). Le Matin is a French daily newspaper published from 1883 to 1944. We built training, validation and test sets by mixing pages from years 1885 and 1937. The ground truth has 9 classes and in the following experiments, we merged the classes related to different levels of titles and also captions with author names. Thus, there are 6 classes.

GBNLA is a dataset given at the competition on German-Brazilian Newspaper Layout Analysis, organized as part of the ICDAR 2019 conference. As the test set has not been shared by the organizers, we split the training set in 3 parts for training, validation and test. The

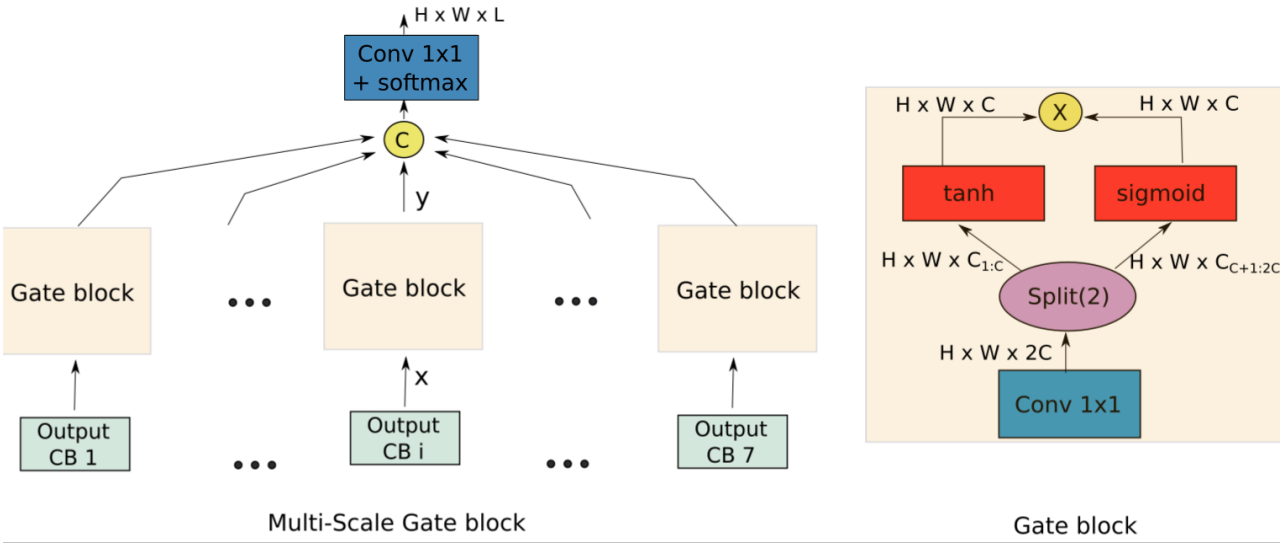

Figure 3: Multi-scale gate block: the outputs of each convolutional block (CB) is given in input of a gate block. H, W and C respectively stands for the height, the width and the channel (or the feature map). The *split(2)* function splits an input tensor of size $H \times W \times 2C$ into two tensors of size $H \times W \times C$. The first tensor feeds a classical tanh activation, while the second controls the gate, modeled using a sigmoid activation. The outputs of gate blocks are concatenated and given as input of the final convolutional layer that makes a prediction by producing L feature maps, where L is the number of labels.

dataset consists of images from 6 historical newspapers and there are 4 classes.

Finally, Luxembourg historical newspapers come from the National Library of Luxembourg (2019). We work on the *ML Starter Pack* containing 1220 pages of the *L'indépendance Luxembourgeoise* (IL) newspaper from 1877, split in 980, 120 and 120 examples for training, validation and test. The second dataset is the *Luxemburger Wort* (LW) newspaper from 1877 split in 76, 12 and 16 examples for training, validation and test.

Images are generally of high resolution. We reduce the input resolution to max 1250 high or 750 wide and keeping the aspect ratio, and apply a sliding window process to extract sub-images of size 600×450 with an overlapping equal to 128 pixels, both in width and height. Sub-images (from 4 to 6 according to the dataset) are standardized and given in input of the network. This requires less memory as the network will be lighter.

4.3. Results

To evaluate our approach, we compute the accuracy from the pixel labeling task. We first compare the 3 network architectures on the 4 historical newspaper datasets. Then, we present a deeper analysis and we show the model performance following the receptive field.

4.3.1. Comparison with standard architectures

Based on the evaluation settings done by Everingham et al. (2015), we evaluate our approach by computing the accuracy (acc) and mean average precision (mAP) from the pixel labeling. Table 3 presents those metrics on each newspapers dataset. The MSGB FCN generally outperforms the two other models on all datasets. While results are slightly better using the MSGB FCN, the system seems more robust than the Dense FCN which obtained lower performance on the IL dataset. While our approach obtains a similar accuracy on the GBNLA dataset than the dense architecture, the mAP is slightly better using the Dense FCN. This may be due to the self-attention mechanism that makes the model overfit when it is trained on a small dataset. With larger datasets, the MSGB FCN benefits from the multi-scale analysis and self-attention mechanism, that focus the decision process on features of interest at various scales, to be more robust than the two

Table 3: Accuracy (acc) and mean average precision (mAP) of pixel labeling provided by the 3 network architectures (Dilated FCN, Dense FCN and MSG FCN described in section 4.1, on the 4 historical newspapers Le Matin (LM), GBNLA (GB) and Luxembourg (IL and LW) datasets.

	Acc				mAP			
	LM	GB	LW	IL	LM	GB	LW	IL
FCN	85.82	81.36	73.26	85.93	87.68	82.12	47.27	53.87
DFCN	86.33	88.29	76.33	90.03	88.22	**84.00**	81.91	83.25
MSGB FCN	**86.49**	**88.31**	**76.53**	**90.79**	**88.44**	83.18	**82.78**	**85.19**

other network architectures both in accuracy and mAP.

4.3.2. Increasing the receptive fields

Finally, we observe the performance of the MSGB FCN depending on the receptive field size. We compare our approach with a state-of-the-art method proposed by Renton et al. (2018). As shown in Table 4, widening the receptive field may strongly improve the recognition, especially with a higher dilation factor. In particular, this is shown on the Luxembourg newspapers, where both the accuracy and mAP are strongly improved by increasing the dilation factor from 2 to 3. The Luxembourg datasets contain some classes relative to small objects in images (e.g. captions or author names) that are poorly represented in the dataset. Thus, having more context to make a decision has a strong positive impact. Increasing the kernel size allows to extract more local context and it may help to improve performance. However, it also increases the number of parameters from $668k$ to $1,585k$ and we can observe on GBNLA dataset that it does not improve performance. This may be due to a lack of training examples and to the high variability of the examples, as there are 7 different newspapers in the dataset.

In addition, Table 4 shows that the FCN proposed by Renton et al. (2018) (FCN 11 R) exhibits poorer performance than our approach. Our approach has less parameters than the FCN 11 R (more than 10 millions vs. at most 1.5 million parameters) but the higher receptive fields and the specific network architecture reach better results. The only dataset where it performs lower is the GBNLA which is a particularly hard dataset as discussed before.

Finally, Figure 4 shows some outputs produced by the MSGB FCN. The MSG block combined with dilated convolutions contributes to predict both homogeneous blocks of various sizes and background between blocks and columns. We can also note that our method may predict multiple labels in a region in case of confusing observations, e.g. text in an advertisement or can sometimes fail when an advertisement is expressed as an article. This is shown on the second top column in images on the left on Fig.4: the image on the bottom left shows that the second block is similar to an article while the label (on the top left image) show that it is an advertisement (a green block). Our model predicted this block as an article with a title and a text block (in purple and red).

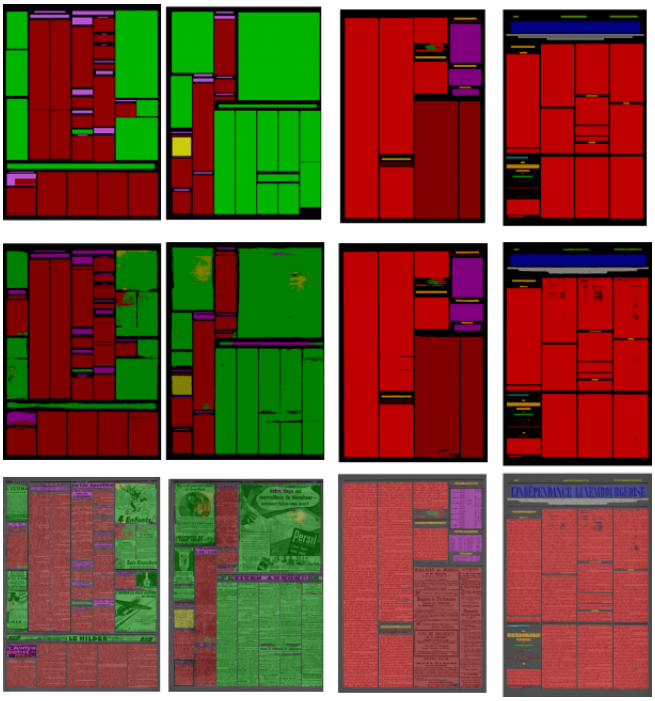

Figure 4: Illustration of pixel prediction. In columns, from the left to the right: two images coming from Le Matin newspaper and two images from Luxembourg newspaper. From top to bottom: ground truth (first line), pixel labeling using the MSGB FCN (second line) and pixel labeling on the original input image (third line).

Table 4: Accuracy (acc) and mean average precision (mAP) on Le Matin and Luxembourg newspapers using the MSGB FCN, following the receptive field size. Various receptive field sizes are evaluated with dilation factor of 2 or 3 and 3x3 or 5x5 kernels. We also compare with the FCN 11 defined by Renton et al. (2018), named FCN 11 R here.

	acc					mAP				
	FCN 11 R	K3 D2	K5 D2	K3 D3	K5 D3	FCN 11 R	K3 D2	K5 D2	K3 D3	K5 D3
LM	78.64	86.49	89.01	**89.23**	87.96	77.50	88.44	**89.79**	89.67	88.29
GB	**91.53**	88.31	86.91	88.21	87.68	**88.53**	83.18	84.37	85.63	82.29
LW	77.16	76.53	78.87	81.35	**90.78**	42.52	82.78	85.50	81.38	**86.49**
IL	86.49	90.79	93.18	94.22	**95.27**	78.93	85.19	89.67	90.59	**90.69**
receptive field (px)	41	89	177	213	425	41	89	177	213	425

5. Conclusion

In this article, a FCN performing region identification by pixel labeling is proposed. The model is based on a multi-scale gate block that benefits of dense connections and a gating mechanism to perform both a multi-scale analysis and self-attention to make a decision. The MSG block faces up the challenging problem of local and contextual analysis required in historical newspaper images to get a semantic labeling.

The proposed model performs well compared to Dilated FCN and Dense FCN, which are state-of-the-art methods for semantic segmentation tasks. The resulting architecture is a very light model (less than 0.7 million of parameters) compared to convolutional models such as VGG and dhSegment Oliveira et al. (2018). This lightweight network overfits less and generalizes more, while requiring less data and speeding up the training process.

This work will be integrated in the PIVAJ software we developed the last few years (Hebert et al. (2014);Tranouez et al. (2015)). It will improve both the pixel level region identification step and the global structure analysis. Most of all, it will streamline the engineered pipeline made of several steps for the same task. However, the PIVAJ software provides a further analysis, including articles identification. The next steps for this work will focus on extracting the structural grid and reading order, in order to build by itself the whole tree of the newspaper issue structure.

Acknowledgments

This work was funded by the French Region Normandy and the European Union as part of the RIN-ASTURIAS grant project, and during Yann Soullard's postdoctoral year at LITIS. Europe acts in Normandy with the European Regional Development Fund (ERDF).

References

Antonacopoulos, A., Clausner, C., Papadopoulos, C., Pletschacher, S., 2013. Icdar 2013 competition on historical newspaper layout analysis (hnla 2013), in: ICDAR, IEEE. pp. 1454–1458.

Bukhari, S.S., Gupta, A., Tiwari, A.K., Dengel, A., 2018. High performance layout analysis of medieval european document images., in: ICPRAM, pp. 324–331.

Chen, K., Seuret, M., Hennebert, J., Ingold, R., 2017. Convolutional neural networks for page segmentation of historical document images, in: ICDAR, IEEE. pp. 965–970.

Coüasnon, B., Lemaitre, A., 2017. Dmos, it's your turn!, in: 1st International Workshop on Open Services and Tools for Document Analysis (ICDAR-OST).

Diem, M., Kleber, F., Fiel, S., Grüning, T., Gatos, B., 2017. cbad: Icdar2017 competition on baseline detection, in: ICDAR, IEEE. pp. 1355–1360.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: A retrospective. International journal of computer vision 111, 98–136.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR, pp. 770–778.

Hebert, D., Palfray, T., Nicolas, S., Tranouez, P., Paquet, T., 2014. Automatic article extraction in old newspapers digitized collections, in: International Conference on Digital Access to Textual Cultural Heritage, ACM. pp. 3–8.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: CVPR, pp. 4700–4708.

Ingle, R.R., Fujii, Y., Deselaers, T., Baccash, J., Popat, A.C., 2019. A scalable handwritten text recognition system. arXiv preprint arXiv:1904.09150 .

Kettunen, K., Ruokolainen, T., Liukkonen, E., Tranouez, P., Antelme, D., Paquet, T., 2019. Detecting articles in a digitized finnish historical newspaper collection 1771-1929: Early results using the pivaj software, in: DATech 2019.

Lemaitre, A., Camillerapp, J., Carton, C., Coüasnon, B., 2018. A combined strategy of analysis for the localization of heterogeneous form fields in ancient pre-printed records. IJDAR 21, 269–282.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: CVPR, pp. 3431–3440.

National Library of Luxembourg, B., 2019. Luxembourg historical newspapers. URL: `https://data.bnl.lu/data/historical-newspapers/`.

Moysset, B., Kermorvant, C., Wolf, C., 2018. Learning to detect, localize and recognize many text objects in document images from few examples. International Journal on Document Analysis and Recognition 21, 161–175.

Oliveira, S.A., Seguin, B., Kaplan, F., 2018. dhsegment: A generic deep-learning approach for document segmentation, in: 2018 16th International Conference on Frontiers in Handwriting Recognition, IEEE. pp. 7–12.

Quirós, L., 2018. Multi-task handwritten document layout analysis. arXiv preprint arXiv:1806.08852 .

Renton, G., Soullard, Y., Chatelain, C., Adam, S., Kermorvant, C., Paquet, T., 2018. Fully convolutional network with dilated convolutions for handwritten text line segmentation. IJDAR 21, 177–186.

Riedl, M., Betz, D., Padó, S., 2019. Clustering-based article identification in historical newspapers, in: Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 12–17.

Sfikas, G., Louloudis, G., Stamatopoulos, N., Gatos, B., 2016. Bayesian mixture models on connected components for newspaper article segmentation, in: ACM Symposium on Document Engineering, pp. 143–146.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks, in: CVPR, pp. 648–656.

Tranouez, P., Nicolas, S., Lerouge, J., Paquet, T., 2015. Pivaj: an article-centered platform for digitized newspapers, in: Archiving Conference, Society for Imaging Science and Technology. pp. 40–43.

Vasilopoulos, N., Kavallieratou, E., 2017. Complex layout analysis based on contour classification and morphological operations. Engineering Applications of Artificial Intelligence 65, 220–229.

Wick, C., Puppe, F., 2018. Fully convolutional neural networks for page segmentation of historical document images, in: IAPR International Workshop on Document Analysis Systems (DAS), IEEE. pp. 287–292.

Xu, Y., Yin, F., Zhang, Z., Liu, C., et al., 2018. Multi-task layout analysis for historical handwritten documents using fully convolutional networks .

Yousef, M., Hussain, K.F., Mohammed, U.S., 2018. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. arXiv preprint arXiv:1812.11894 .

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 .