DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction

Akansha Bhardwaj[1,2] , Dominik Mercier[1] , Sheraz Ahmed[1] , Andreas Dengel[1]

[1] Smart Data and Services, DFKI Kaiserslautern, Germany
firstname.lastname@dfki.de
[2] eXascale Infolab, University of Fribourg, Switzerland
akbwaj@exascale.info

Abstract. Extraction of structured bibliographic data from document images of non-native-digital academic content is a challenging problem that finds its application in the automation of cataloging systems in libraries and reference linking domain. The existing approaches discard the visual cues and focus on converting the document image to text and further identifying citation strings using trained segmentation models. Apart from the large training data, which these existing methods require, they are also language dependent. This paper presents a novel approach (DeepBIBX) which targets this problem from a computer vision perspective and uses deep learning to semantically segment the individual citation strings in a document image. DeepBIBX is based on deep Fully Convolutional Networks and uses transfer learning to extract bibliographic references from document images. Unlike existing approaches which use textual content to semantically segment bibliographic references, DeepBIBX utilizes image based contextual information, which makes it applicable to documents of any language. To gauge the performance of the presented approach, a dataset consisting of 286 document images containing 5090 bibliographic references is collected. Evaluation results reveals that the DeepBIBX outperforms state-of-the-art method (ParsCit, 71.7%) for bibliographic references extraction and achieved an accuracy of 84.9% in comparison to 71.7%. Furthermore, in terms of pixel classification task, DeepBIBX achieved a precision and a recall rate of 96.2%, 94.4% respectively.

Keywords: Deep learning, Machine learning, Bibliographic data, Reference linking

1 Introduction

The delivery of knowledge through the digital format has enabled readers to access and share knowledge around the world. This phenomenon has resulted in digital becoming the regular format, owing to the ease of accessing, preserving and sharing content. Though digital content is ubiquitous, the content which is not natively-digital continues to lack suitable metadata which makes it difficult for such content to be easily discoverable. The most obvious example

of such content is the digitization of old articles, where the scanning process renders a digital image. For the successful implementation of digital libraries, it is important to automate the generation of bibliographic databases for non natively-digital books.

Most of the works done so far on the task of generating bibliographic databases focus on converting the image into text and then further using the text segmentation techniques to structure citation data [1],[2]. This results in the loss of contextual information present in bibliographic document images which has the discriminative ability to identify references from one another. In this work, we introduce an image based reference extraction model where the above problem has been approached from a deep learning perspective.

Deep learning has recently proven to be extremely successful on various tasks of visual recognition [3],[4],[5] including semantic segmentation [6]. In this work, we introduce a semantic segmentation model for image based reference extraction. The issue of unavailability of large amount of training data for this model has been bypassed using the transfer learning approach introduced by Carauana [7]. Also, transfer learning benefits by saving the additional cost of time and computational resources needed to perform training on a large scale.

In this work, we have transferred the knowledge gained from the FCN-8s network [6] trained for PASCAL VOC challenge [8] with 21 classes to identifying individual citation strings based on the contextual indentation information present in bibliographic document images.

1.1 Paper Contribution

This paper introduces a novel deep learning based semantic segmentation model fine-tuned for reference extraction in bibliographic document images. The trained model detects individual citation strings in a document image with a precision of 83.9% and a recall of 84.6%. The work also presents a framework in which references are identified in a document image, converted to text and further resolved to structured segmented information for reference linking applications. Further evaluations with the state-of-the-art ParsCit segmentation model [1] show that while ParsCit extracted 71.7% citation strings, our approach extracts 84.9% citation strings on a test set of 286 bibliographic document images. This approach of identifying individual citation strings works for bibliographic document images of any language as it utilizes the contextual indentation information present in document images.

2 Related Works

Procedures for digitization of books have improved considerably in past few years and several ambitious projects with libraries aim to digitize thousands of books. Google book-scanning project [9] works with libraries to offer digitized books and aims to digitize every book ever printed.

In this phase of transition, digital libraries require automated cataloging process for the purpose of generating bibliographic databases. Several projects have focused on the creation of an electronic card catalog and the success of these endeavors resulted in Online Public Access Catalog (OPAC) replacing the traditional card catalog in many academic, public and special libraries. Bibliographic data has more power now with new technologies assisting it's reuse in research with citation management softwares, linking of data from multiple sources and also in aiding data mining of large datatsets to identify publication trends.

Several projects have focused on automatic identification of references in scholarly PDF documents but, they either rely only on the text [1],[2], or employ heuristics based on visual cues [10].

To the best of our knowledge, none of the approaches so far have used deep learning to identify bounding boxes of citation strings in document images.

3 Dataset

The data for this task has been collected from print media (books, journals, articles, etc) present in libraries arranged in different indentation formats. The bibliographic information belonging to these print media have been scanned at a DPI resolution of 300 or more. For creation of the dataset, an equal number of files are selected from each category of publication to generate a well-balanced dataset of 440 files. These files are manually annotated with bounding boxes around each reference. Fig. 1 shows some of the sample data.

This data is further augmented by removing the whitespace around the text area and the final transformed dataset consisted of 574 train images, 50 images for validation and 298 images for test set. All images are further cropped to a $width \times height$ of 500×1500 pixels. The labels of the dataset consist of a rectangular bounding box surrounding each reference, the rest is labeled as background.

4 DeepBIBX: The Proposed Approach

Extraction of structured data from bibliographic document images starts with pre-processing bibliographic document images, followed by deep learning based

(a) (b) (c) (d)

Fig. 1. Examples of various kinds of bibliographic document images in raw dataset

reference extraction from document images, and further segmentation of each extracted citation string into structured information like title, author, publisher, volume, pages, etc.

4.1 Preprocessing

Before starting the actual pipeline of reference extraction, it is important to normalize scanned document images for different qualitative distortions. To do so, document binarization is performed to convert color and gray scale documents into binary format. The method described by Breuel [11] is used to perform document binarization and to correct document skew. The binarized and skew corrected document image is then passed to the heart of DeepBIBX, i.e., Image based reference extraction module.

4.2 Image based Reference Extraction

Architecture Fig. 2 shows the architecture of a deep Fully Convolutional Network (FCN) [6] which is used for semantic segmentation. FCNs take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. Each layer of data is a three-dimensional array of size $h \times w \times d$, where h and w are $height, width$ respectively, and d is the color channel dimension. Receptive fields are the locations in higher layers which are connected to the locations in the image.

The basic components of an FCN consists of convolution, pooling, and activation functions. They operate on local input regions, and depend only on relative spatial coordinates. If x_{ij} is the data vector at location $(i, j)$for a particular layer, and y_{ij} is the data vector for the following layer, these functions compute outputs y_{ij} by

$$y_{ij} = f_{ks}(\{x_{si+\delta i, sj+\delta j}\}_{0<=\delta i, \delta j<=k})$$

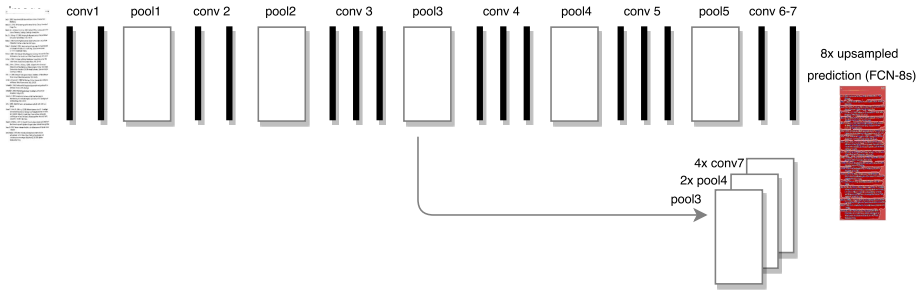


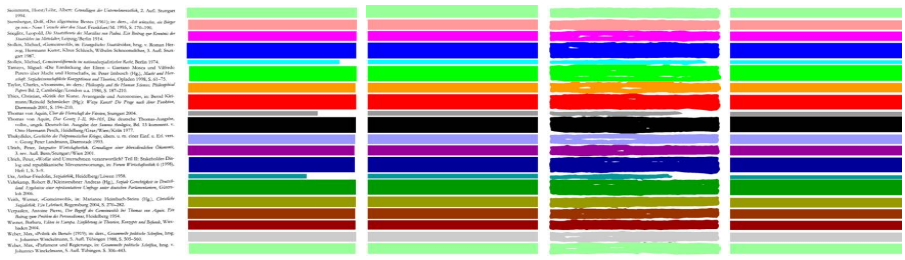
Fig. 2. FCN-8s architecture: Pooling and prediction layers are shown as boxes while intermediate layers are shown as vertical lines. Additional predictions from pool3, at stride 8, provide further precision

where k is called the kernel size, s is the stride or subsampling factor, and f_{ks} determines the layer type: a matrix multiplication for convolution or average pooling, a spatial max for max pooling, or an element nonlinearity for an activation function. An FCN operates on an input of any size, and produces an output of corresponding spatial dimensions.

A real-valued loss function composed with an FCN defines a task. If the loss function is a sum over the spatial dimensions of the final layer, its gradient will be a sum over the gradients of each of its spatial components, considering all of the final layer receptive fields as a minibatch.

Approach Deep learning based approaches usually require a lot of training data. However, a large amount of data for our task was unavailable. To resolve this problem, the concept of transfer learning was adopted. In DeepBIBX, FCN-8 network, which was pre-trained on PASCAL VOC 21 class challenge [8] problem, was used to allow for better segmentation for reference extraction. The last layer of original FCN-8 network which in its default settings outputs 21 classes is removed and the activations of the last hidden layer are used as the feature descriptors of the input dataset. A final layer is added which outputs 2 classes, reference area and background.

The network is trained on the dataset for 80 epochs with stochastic gradient descent. The semantic segmentation output generated after training separates the foreground from background roughly (refer Fig. 3d). This output is further post-processed to obtain crisper boundaries using blob identification heuristics and the output is transformed as shown in Fig. 4e. It is important to mention here that these heuristics have been developed only on the validation set. Each bounding box in the post-processed result is compared to each bounding box present in the ground truth. A box is identified if the Intersection over Union (IoU) ratio for resulting bounding box and processed ground truth bounding box



(a) (b) (c) (d) (e)

Fig. 3. An example from test set where each detected box has an IoU>=0.5 resulting in 100% precision and 100% recall (a) Bibliographic document image (b) Corresponding human annotated bounding boxes, (c) processed human annotated data for evaluation (d) prediction generated by semantic segmentation model (3) identification of bounding boxes after post-processing, each labeled with a different color

is greater than 0.5. For each document image, these identified bounding boxes are used to calculate precision and recall. Fig. 3 shows an example of an image from test set where inferred precision and recall rate is 100%.

In the next experiment, all text lines of the ground truth document image have been replaced with rectangular boxes, the same length and width as that of each text line. Keeping other parameters same as in previous experiment, pre-trained FCN-8s was trained on this transformed dataset as well. Figure 4 shows visualization of the activation intensities at multiple layers during the forward pass of a test image.

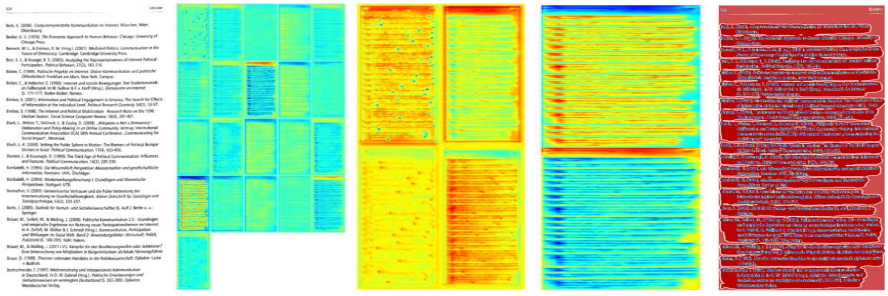
4.3 Segmentation of citation strings

Once the region of each individual bibliographic entry is identified, OCR is performed for each individual entry [11]. The resulting textual information can be given as an input to any citation string segmentation model like AnyStyle [12], ParsCit [1] to segment it into author, title, DOI, page, publisher, volume and other relevant information which can be used for reference linking applications.

5 Evaluation

Due to varying parameters of the model, a complete match between predicted and ground-truth bounding boxes is unrealistic. Therefore, approaches based on semantic segmentation are evaluated using the IoU metric. This metric rewards predicted bounding boxes for heavily overlapping with the ground-truth bounding boxes.

The pixel-wise evaluation for bounding box on document image gives a precision of 96.2% and a recall of 94.4%. It is important to mention here that though the pixel wise evaluation results in good precision and recall, it is not a good



(a) (b) (c) (d) (e)

Fig. 4. Visualization of the activation intensities of FCN-8s network at multiple layers during the forward pass of a test image. (a) Source image (b) fuse_pool3 (c) upscore8 (d) score_2 classes (e) inference image

evaluation measure as the trained model might simply be a text line recognizer. To make sure that this is not the case, results have been further evaluated for detection of each bounding box.

For further evaluation, a reference box is identified when the IoU of predicted bounding box and labeled bounding box is greater than 0.5 after post-processing. This results in a precision of 82.6% and a recall of 80.0%. In the experiment, where the lines were blurred, a precision of 83.9% and a recall of 84.6% was observed. Fig. 5 shows a precision and recall curve with respect to IoU for the case when text lines are blurred.

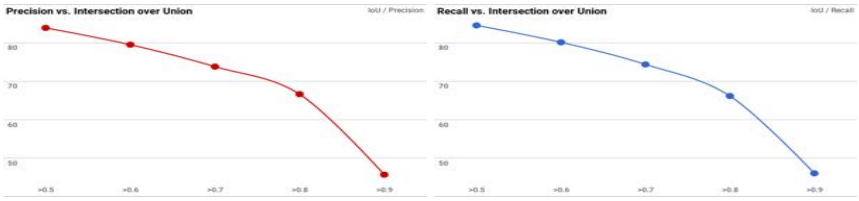
Table 1. Evaluation of image based extraction results

Category	Precision	Recall
Pixel-wise evaluation	96.2%	94.4%
Bounding box detection on plain document image	82.6%	80.0%
Bounding box detection on document image with blurred lines	83.9%	84.6%

ParsCit is the current state of the art in the area of reference segmentation. Table 2 compares the results from our approach to ParsCit. The results show that on a test set of 286 bibliographic document images, which were converted to text, ParsCit extracted 3645 references and our image based reference extraction model extracted 4323 references out of a total of 5090 references. These results suggest that visual cues are very important during identification of references and should not be discarded.

Table 2. Evaluation of results when compared to ParsCit

Category	Number of extracted references	Extracted percentage
ParsCit	3645	71.7%
Proposed approach	4323	84.9%



(a) (b)

Fig. 5. Precision, Recall results compared to varying IoU (a) Precision vs. IoU (b) Recall vs. IoU

This approach of image based reference extraction with an OCR tool and a segmentation tool [1] can be used for automation of cataloging systems in libraries or for reference linking applications.

6 Conclusion and Future Work

This work presents a novel deep learning based semantic segmentation model for identifying references in bibliographical document images. This model is language independent and identifies individual references with a precision of 83.9% and a recall rate of 84.6%. The results have been compared with state-of-the-art text based semantic segmentation model ParsCit where the proposed model outperforms the reference detection task by a margin of more than 13%. These results suggest that utilizing the contextual information present in bibliographic document images is a key factor in extraction of bibliographic data.

This work is useful for the automation of library cataloging systems and for reference linking applications. The future work will focus on a comprehensive model for the above tasks and provide a solution for digital libraries.

References

1. Councill, Isaac G., C. Lee Giles, and Min-Yen Kan. "ParsCit: an Open-source CRF Reference String Parsing Package." LREC. Vol. 2008. 2008.
2. Tkaczyk, Dominika, et al. "CERMINE: automatic extraction of structured metadata from scientific literature." International Journal on Document Analysis and Recognition (IJDAR) 18.4 (2015): 317-335.
3. K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385, 2015.
4. X. Zhang, Z. Li, C. C. Loy, and D. Lin, Polynet: A pursuit of structural diversity in very deep networks, arXiv preprint arXiv:1611.05725, 2016
5. C. Szegedy, S. Ioffe, and V. Vanhoucke, Inception-v4, inception-resnet and the impact of residual connections on learning, arXiv preprint arXiv:1602.07261, 2016.
6. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
7. Caruana, Multitask learning, in Learning to learn. Springer, 1998, pp. 95133
8. Everingham, M., et al. "The PASCAL visual object classes challenge 2012 results." See http://www. pascal-network. org/challenges/VOC/voc2012/workshop/index. html. Vol. 5. 2012.
9. Committee on Institutional Cooperation: Partnership announced between CIC and Google, 6 June 2007, Retrieved 7.
10. Crossref.org. (2017). pdfextract - Crossref. [online] Available at: https://www.crossref.org/labs/pdfextract/ [Accessed 20 Aug. 2017].
11. Breuel, Thomas M. "The OCRopus open source OCR system." DRR 6815 (2008): 68150.
12. Anystyle.io. (2017). AnyStyle.io. [online] Available at: https://anystyle.io/ [Accessed 20 Aug. 2017].