An Approach How to Automate Labeling Data for the Training ANN Models for Page Layout Analysis

Andrey Mikhailov[1]

Matrosov Institute for System Dynamics and Control Theory of SB RAS,
134 Lermontov st., Irkutsk, Russia
mikhailov@icc.ru,
WWW home page: http://idstu.irk.ru

Abstract. Object detection and recognition is an important task in many document analysis applications. It is a difficult problem due to different page layouts and representation formats. Recently the deep learning in computer vision has significantly boosted the data-driven image-based approaches for page layout analysis. In this paper, we consider open formats of electronic documents to generate training datasets. Formats of these documents should contain markup allowing obtaining information about page layout regions. It will allow us to generate a training dataset automatically for training ANN models of page layout analysis.

Keywords: document layout analysis · PDF accessibility · ANN models · artificial intelligence

1 Introduction

Arbitrary documents are a common way of presenting information on the web. The big volume and structure of such documents make them a valuable source in data science and business intelligence applications. However, as a rule, they haven't included semantics for machine interpretation of their content as considered by their author. The information accumulated in them is often unstructured and not standardized. The analysis of these data requires transformation to a structured representation with a given formal model. In document analysis and recognition, this task commonly named as document layout analysis. In recent years, approaches for page layout analysis based on deep neural networks for object detection and classification have been actively developing. This is evidenced by the results of one of the main scientific conferences on document analysis - ICDAR [1]. Since 2001, this conference has hosted the RDCL document

[1] https://icdar2019.org

layout analysis competitions (2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019). For the competition are published datasets. For example, in 2019 a dataset was published that includes only 4, 500 examples.

This amount of data is not enough for high-quality training of deep neural networks, since their modern architectures have many free parameters and are very sensitive to the volume and quality of data. Modern layout analysis systems based on ANN models are focused mainly on a small count of the same document types. This is due to the fact that either open-source or hand-tagged datasets were used to develop page layout analysis ANN models. In this paper, we propose an idea to automate the process of labeling datasets. For this, it is proposed to develop methods for automatic data labeling for training deep neural for page layout analysis. Which should reduce the process of developing layout analysis systems for new types of documents, and improve the quality of the analysis.

2 Related Works

Document images are often generated from physical documents by digitization, using scanners or various generation programs (printers). Many documents, such as newspapers, magazines and brochures, have very complex layouts due to the placement of pictures, headings and captions, complex backgrounds, artistic text formatting, etc.

A person uses a lot of additional clues such as context, conventions, language information. Automatic analysis of an arbitrary document with a complex layout is an extremely difficult task and goes beyond the capabilities of modern document layout analysis systems. In the scientific literature, a large number of methods for analyzing the layout of documents have been proposed. According to article [10], they can be divided into three groups: methods of classification based on areas [17, 13]; classification methods based on pixel analysis [12, 11]; analysis of connected components [6, 15, 3]. With the increasing efficiency and popularity of convolutional neural networks, their field of application is constantly expanding. Since 2014, the first attempts to use artificial neural networks to solve the problem of analyzing the layout of documents have been known [9, 8, 2, 16]. These works have demonstrated their effectiveness in comparison with classical approaches, which is confirmed by the results of the 2017 competition at the ICDAR conference [4]. On the other hand, the 2019 competition showed that on a variety of data, with a large number of classes (10), the combination of classical methods [5] is most effective compared to deep neural networks. This is due to the lack of a sufficient amount of diverse tagged data with a large number of classes. While for special cases, neural networks work much more efficiently [7]. It should be noted that to solve the problem of analyzing the arrangement of documents in these works, either neural networks of the R-CNN architecture or author's developments are used. For training neural networks, open datasets of labeled data are usually used; in rare cases, the authors of the articles indicate that they have labeled their own training set. These samples rarely reach 20,000 copies and are often not publicly available. The author is not aware at the mo-

ment of open datasets large enough to train neural networks for document layout analysis. It should be noted that it was the creation of such datasets as ImageNet that made it possible to obtain outstanding results using convolutional neural networks for natural image recognition.

3 An Idea

The Internet contains a large number of original LaTex documents. One of the most well-known resources is arXiv [2]. arXiv is a free distribution service and an open-access archive for more than 1,7 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. PDF documents can be generated from these documents using special compiler programs. PDF is a page orientated graphic format. It simply puts images and glyphs at various coordinates on a page.

Since 2006, PDF includes special tags for support reading order and logical order. With reading order, the characters on the page are understood to have a linear sequence of appearance. Logical order allows introducing concepts such as tables, lists, and headings, as well as provide alternate text for images, descriptive text for links and form fields, and so on.

Traditionally, there are three ways to obtain PDF document from LaTex.

- LaTeX source file converted to a DVI file, which could then be converted to PostScript with dvips. This, in turn, can be converted to a PDF file by ps2pdf [3] tool.

```
          latex             dvips            ps2pdf
text.tex -------> text.dvi -------> text.ps --------> text.pdf
```

- The step with conversion to PostScript can be skipped.

```
          latex           dvipdfm
text.tex -------> text.dvi -------> text.pdf
```

- Directly from the LaTeX source to PDF file by pdflatex program.

```
         pdflatex
text.tex --------> text.pdf
```

The first two ways are not allows for tagging PDF. Because the DVI format does not allow saving additional tags. For the direct compilation of LaTex into PDF there is a special LaTex package - accessibility [4].

[2] https://arxiv.org/
[3] https://www.ps2pdf.com/
[4] https://github.com/AndyClifton/accessibility

Accessibility [14] was written as a proof-of-concept showing how to improve the structure and tagging of PDF files generated from LaTeX. These features make PDF documents machine-readable and thus enable document readers to automatically process and present the document. Andy Clifton took on maintenance of the package in May 2019 with permission and support from Babett Schalitz. This package is predominantly targeted at documents produced using the KOMA-Script document classes [1].

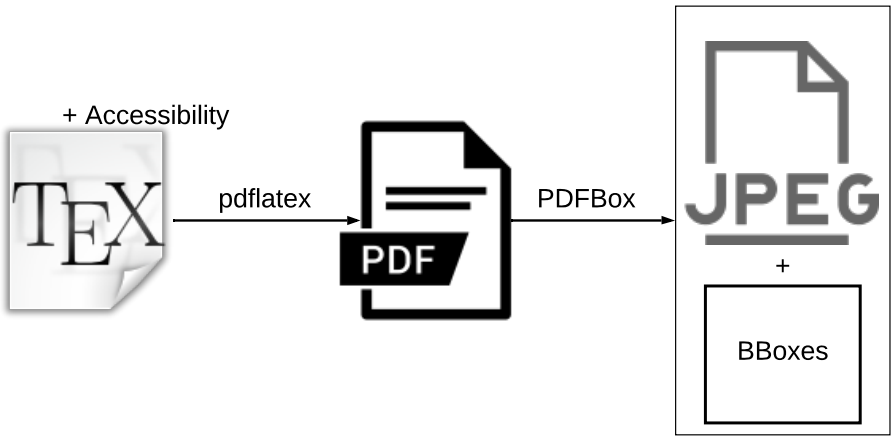

Fig. 1. Labeling PDFs process

The idea of automating data labeling is shown in the figure [**img˙concept**]. We propose to use the Accessibility package for generating tagged PDF documents. This package is easy to use. In order to get a tagged document, the only short preamble is needed to add to the document and compiled using pdflatex tool. The next step is to extract the tagged information from the tagged document. PDFBox allows to extract content from documents and render PDF to image. We propose to use this tool to generate training dataset from tagged PDFs.

4 Conclusion

In this paper, we presented an idea how to automate dataset labeling from LaTex documents. The main idea is use special LaTex package Accessibility. This package allows adding tags to produced PDF documents. To extract information about layout from tagged PDFs we suggest to use PDFBox library. We expect that the explained principles can be used for designing software for page layout analysis.

Acknowledgment

The research was supported by the Program of the Fundamental Research of the Siberian Branch of the Russian Academy of Sciences, project num. IV.38.1.2 (reg. num. AAAA-A17-117032210079-1). Results are achieved using the Centre of collective usage Integrated information network of Irkutsk scientific educational complex.

References

[1] AndyClifton. *The Accessibility LaTeX package*. 2019 (accessed August 21, 2020). URL: `https://github.com/AndyClifton/accessibility`.

[2] Dario Augusto Borges Oliveira and Matheus Palhares Viana. "Fast CNN-based document layout analysis". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 1173–1180.

[3] Syed Saqib Bukhari, Mayce Ibrahim Ali Al Azawi, Faisal Shafait, and Thomas M Breuel. "Document image segmentation using discriminative learning over connected components". In: *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. 2010, pp. 183–190.

[4] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. "Icdar2017 competition on recognition of documents with complex layouts-rdcl2017". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 1404–1410.

[5] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. "ICDAR2019 competition on recognition of documents with complex layouts-rdcl2019". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1521–1526.

[6] Lloyd A. Fletcher and Rangachar Kasturi. "A robust algorithm for text string separation from mixed text/graphics images". In: *IEEE transactions on pattern analysis and machine intelligence* 10.6 (1988), pp. 910–918.

[7] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. "Icdar 2019 competition on table detection and recognition (ctdar)". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1510–1515.

[8] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. "Evaluation of deep convolutional nets for document image classification and retrieval". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2015, pp. 991–995.

[9] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. "Convolutional neural networks for document image classification". In: *2014 22nd International Conference on Pattern Recognition*. IEEE. 2014, pp. 3168–3172.

[10] Viet Phuong Le, Nibal Nayef, Muriel Visani, Jean-Marc Ogier, and Cao De Tran. "Text and non-text segmentation based on connected component features". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2015, pp. 1096–1100.

[11] Michael A Moll and Henry S Baird. "Segmentation-based retrieval of document images from diverse collections". In: *Document Recognition and Retrieval XV*. Vol. 6815. International Society for Optics and Photonics. 2008, p. 68150L.

[12] Michael A Moll, Henry S Baird, and Chang An. "Truthing for pixel-accurate segmentation". In: *2008 The Eighth IAPR International Workshop on Document Analysis Systems*. IEEE. 2008, pp. 379–385.

[13] Oleg Okun, David Dœrmann, and Matti Pietikainen. *Page segmentation and zone classification: the state of the art*. Tech. rep. OULU UNIV (FINLAND) DEPT OF ELECTRICAL ENGINEERING, 1999.

[14] Babett Schalitz. *Accessibility-erhöhung von latex-dokumenten. Diplomarbeit, Fakultät Informatik*. Tech. rep. Technische Universität Dresden, July 2007.

[15] Karl Tombre, Salvatore Tabbone, Loıc Pélissier, Bart Lamiroy, and Philippe Dosch. "Text/graphics separation revisited". In: *International Workshop on Document Analysis Systems*. Springer. 2002, pp. 200–211.

[16] Nicole Vincent and Jean-Marc Ogier. "Shall deep learning be the mandatory future of document analysis problems?" In: *Pattern Recognition* 86 (2019), pp. 281–289.

[17] Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl. "Document analysis system". In: *IBM journal of research and development* 26.6 (1982), pp. 647–656.