DocBank: A Benchmark Dataset for Document Layout Analysis

Minghao Li[1*], **Yiheng Xu**[2*], **Lei Cui**[2], **Shaohan Huang**[2],
Furu Wei[2], **Zhoujun Li**[1], **Ming Zhou**[2]
[1]Beihang University
[2]Microsoft Research Asia
{liminghao1630,lizj}@buaa.edu.cn
{v-yixu,lecu,shaohanh,fuwei,mingzhou}@microsoft.com

Abstract

Document layout analysis usually relies on computer vision models to understand documents while ignoring textual information that is vital to capture. Meanwhile, high quality labeled datasets with both visual and textual information are still insufficient. In this paper, we present **DocBank**, a benchmark dataset that contains 500K document pages with fine-grained token-level annotations for document layout analysis. DocBank is constructed using a simple yet effective way with weak supervision from the LaTeX documents available on the arXiv.com. With DocBank, models from different modalities can be compared fairly and multi-modal approaches will be further investigated and boost the performance of document layout analysis. We build several strong baselines and manually split train/dev/test sets for evaluation. Experiment results show that models trained on DocBank accurately recognize the layout information for a variety of documents. The DocBank dataset is publicly available at https://github.com/doc-analysis/DocBank.

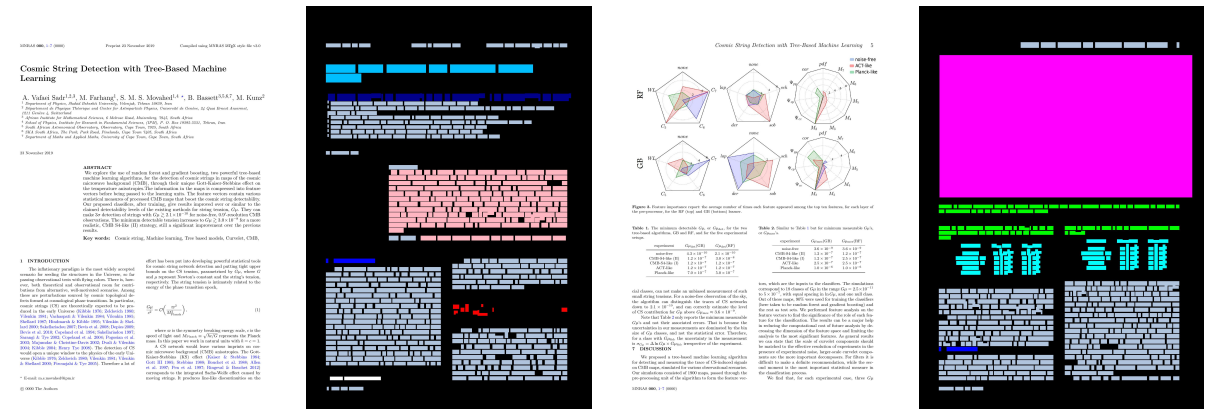
1 Introduction

Document layout analysis is an important task in many document understanding applications as it can transform semi-structured information into a structured representation, meanwhile extracting key information from the documents. It is a challenging problem due to the varying layouts and formats of the documents. Existing techniques have been proposed based on conventional rule-based or machine learning methods, where most of them fail to generalize well because they rely on hand crafted features that may be not robust to layout variations. Recently, the rapid development of deep learning in computer vision has significantly boosted the data-driven image-based approaches for document layout analysis. Although these approaches have been widely adopted and made significant progress, they usually leverage visual features while neglecting textual features from the documents. Therefore, it is inevitable to explore how to leverage the visual and textual information in a unified way for document layout analysis.

Nowadays, the state-of-the-art computer vision and NLP models are often built upon the pre-trained models (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018; Lample and Conneau, 2019; Yang et al., 2019; Dong et al., 2019; Raffel et al., 2019; Xu et al., 2019) followed by fine-tuning on specific downstream tasks, which achieves very promising results. However, pre-trained models not only require large-scale unlabeled data for self-supervised learning, but also need high quality labeled data for task-specific fine-tuning to achieve good performance. For document layout analysis tasks, there have been some image-based document layout datasets, while most of them are built for computer vision approaches and they are difficult to apply to NLP methods. In addition, image-based datasets mainly include the page images and the bounding boxes of large semantic structures, which are not fine-grained token-level annotations. Moreover, it is also time-consuming and labor-intensive to produce human-labeled and fine-grained token-level text block arrangement. Therefore, it is vital to leverage weak

[*]Equal contributions during internship at Microsoft Research Asia.

Proceedings of the 28th International Conference on Computational Linguistics, pages 949–960
Barcelona, Spain (Online), December 8-13, 2020



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

Figure 1: Example annotations of the DocBank. The colors of semantic structure labels are: Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, Title

supervision to obtain fine-grained labeled documents with minimum efforts, meanwhile making the data be easily applied to any NLP and computer vision approaches.

To this end, we build the DocBank dataset, a document-level benchmark that contains 500K document pages with fine-grained token-level annotations for layout analysis. Distinct from the conventional human-labeled datasets, our approach obtains high quality annotations in a simple yet effective way with weak supervision. Inspired by existing document layout annotations (Siegel et al., 2018; Li et al., 2019; Zhong et al., 2019), there are a great number of digital-born documents such as the PDFs of research papers that are compiled by LaTeX using their source code. The LaTeX system contains the explicit semantic structure information using mark-up tags as the building blocks, such as abstract, author, caption, equation, figure, footer, list, paragraph, reference, section, table and title. To distinguish individual semantic structures, we manipulate the source code to specify different colors to the text of different semantic units. In this way, different text zones can be clearly segmented and identified as separate logical roles, which is shown in Figure 1. The advantage of DocBank is that, it can be used in any sequence labeling models from the NLP perspective. Meanwhile, DocBank can also be easily converted into image-based annotations to support object detection models in computer vision. In this way, models from different modalities can be compared fairly using DocBank, and multi-modal approaches will be further investigated and boost the performance of document layout analysis. To verify the effectiveness of DocBank, we conduct experiments using four baseline models: 1) BERT (Devlin et al., 2018), a pre-trained model using only textual information based on the Transformer architecture. 2) RoBERTa (Liu et al., 2019), a robustly optimized method for pre-training the Transformer architecture. 3) LayoutLM (Xu et al., 2019), a multi-modal architecture that integrates both the text information and layout information. 4) Faster R-CNN (Ren et al., 2015), a high performance object detection networks depending on region proposal algorithms to hypothesize object locations. The experiment results show that the LayoutLM model significantly outperforms the BERT and RoBERTa models and the object detection model on DocBank for document layout analysis. We hope DocBank will empower more document layout analysis models, meanwhile promoting more customized network structures to make substantial advances in this area.

The contributions of this paper are summarized as follows:

- We present DocBank, a large-scale dataset that is constructed using a weak supervision approach. It enables models to integrate both the textual and layout information for downstream tasks.
- We conduct a set of experiments with different baseline models and parameter settings, which confirms the effectiveness of DocBank for document layout analysis.
- The DocBank dataset is available at `https://github.com/doc-analysis/DocBank`.

Figure 2: Data processing pipeline

2 Task Definition

The document layout analysis task is to extract the pre-defined semantic units in visually rich documents. Specifically, given a document $\mathcal{D}$ composed of discrete token set $t = \{t_0, t_1, ..., t_n\}$, each token $t_i = (w, (x_0, y_0, x_1, y_1))$ consists of word w and its bounding box (x_0, y_0, x_1, y_1). And $\mathcal{C} = \{c_0, c_1, .., c_m\}$ defines the semantic categories that the tokens are classified into. We intend to find a function $F : (\mathcal{C}, \mathcal{D}) \rightarrow \mathcal{S}$, where $\mathcal{S}$ is the prediction set:

$$\mathcal{S} = \{(\{t_0^0, ..., t_0^{n_0}\}, c_0), ..., (\{t_k^0, ..., t_k^{n_k}\}, c_k)\} \tag{1}$$

3 DocBank

We build DocBank with token-level annotations that supports both NLP and computer vision models. As shown in Figure 2, the construction of DocBank has three steps: Document Acquisition, Semantic Structures Detection, Token Annotation. Meanwhile, DocBank can be converted to the format that is used by computer vision models in a few steps. The current DocBank dataset totally includes 500K document pages, where the training set includes 400K document pages and both the validation set and the test set include 50K document pages.

3.1 Document Acquisition

We download the PDF files on arXiv.com as well as the LaTeX source files since we need to modify the source code to detect the semantic structures. The papers contain Physics, Mathematics, Computer Science and many other areas, which is beneficial for the diversity of DocBank to produce robust models. We focus on English documents in this work and will expand to other languages in the future.

3.2 Semantic Structures Detection

DocBank is a natural extension of the TableBank dataset (Li et al., 2019), where other semantic units are also included for document layout analysis. In this work, the following semantic structures are annotated in DocBank: {Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table and Title}. In TableBank, the tables are labeled with the help of the 'fcolorbox' command. However, for DocBank, the target structures are mainly composed of text, where the 'fcolorbox' cannot be well applied. Therefore, we use the 'color' command to distinguish these semantic structures by changing their font colors into structure-specific colors. Basically, there are two types of commands to represent semantic structures. Some of the LaTeX commands are simple words preceded by a backslash. For instance, the section titles in LaTeX documents are usually in the format as follows:

```
\section{The title of this section}
```

Other commands often start an environment. For instance, the list declaration in LaTeX documents is shown as follows:

```
\begin{itemize}
    \item First item
    \item Second item
\end{itemize}
```

The command $\backslash begin\{itemize\}$ starts an environment while the command $\backslash end\{itemize\}$ ends that environment. The real command name is declared as the parameters of the 'begin' command and the 'end' command.