Deep Table Detection Using Image-Lexical Features

Hyebin Kwon, Joungbin An, and Won-Yong Shin

Yonsei Univeristy

{wdcd6211, jbistanbul05, wy.shin}@yonsei.ac.kr

Abstract

Using the state-of-the-art object detection model, so-called YOLOv5, and custom semantic feature functions, we present a new deep learning-based table detection method that encompasses both image and lexical features to enhance the performance of table detection (in PDF documents). It is demonstrated that the proposed method outperforms YOLOv5 in terms of the F1 score.

1 Introduction

Despite its growing significance, table detection still remains a relatively unresolved problem due to the inconsistency of its form. The form and logic may differ immensely among fields, documents, cultures, and even languages, making it harder to build a universal detection model for tables. Many of earlier studies attempted to tackle the issue through hand-crafted textual heuristics or statistical machine learning methods (mostly natural language processing (NLP) models, such as naïve Bayes, hidden Markov models, or conditional random fields) [1]. While some presented notable results, most models were inefficiently designed and heavily reliant on datasets. Since around 2010, with the advent of deep neural networks, methods have shifted towards utilizing convolutional neural networks (CNNs) to detect tables as an object of an image and have brought forth more promising results [2]. However, in taking the input as an image format, all textual context, which still carries crucial clues, is lost. In testing the CNN-based models, we observed an overall high false positive (FP) rate with models mistaking graphs, mathematical equations, or even plain text as tables. In doing so, it was also noticed that many of these FPs could be eliminated if some lexical information was given before classification. In this paper, we propose a method that encompasses both image features from YOLOv5, a state-of-the-art object detector, and textual information (i.e., lexical features) to successfully detect tables.

2 Methodology

In order to match the format of input for the object detector YOLOv5, all pages of the PDF documents in the dataset are converted to images to pass through the network model. At the same time, the PDF document is passed through two semantic (lexical) feature functions, which capture information that is lost through the process of converting the file to an image.

The first function calculates the number of consecutive irregular spaces of text in lines for each page. While the structure of tables may vary, the fact that tables are spaced differently from plain text is a universal characteristic applicable to any form. Moreover, it excludes other non-plain text forms that are frequently mistaken for tables such as graphs, figures, or mathematical equations.

The next function finds the caption, written as “Table”, “table” or “TABLE”, in a window of 7 lines up and down from the detected irregular spaces. The function is crafted so that it differentiates captions and the word used in plain text for sake of explanation. This feature serves as a reinforcement for determining the existence of a table after finding irregular spaces.

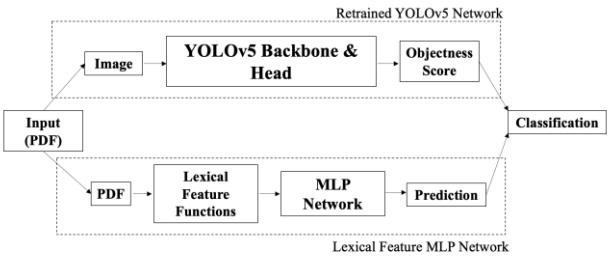

Figure 1. Model Pipeline

With the newly computed lexical features, the outputs are passed through a multi-layer perceptron (MLP) module for classification, outputting the probability of the existence of a table in the input file. This score, along with the objectenss score (also a probability) from the YOLOv5 model is jointly thresholded for classification. The model pipeline is illustrated in Figure 1.

3 Experimental Results

3-1 Experimental Setup

With training and re-training YOLOv5 on the benchmark dataset for table detection ICDAR2013

and a set of Materials Science papers, respectively, we compare the two results of classification with and without the information of lexical features.

The YOLOv5 network was pre-trained on the ICDAR2013 Dataset, which consists of 238 pdf pages. Retraining and testing was done with an 8 to 2 split on the set of Materials Science papers, containing 223 and 56 pdf pages respectively. The experimental results presented below are from tests performed on the test batch of Materials Science papers.

3-2 Experimental Results

Table 1. Comparison in terms of F1 scores
"Baseline" refers to YOLOv5 without the lexical features, and "Proposed" refers to our model, which includes the textual information.
"Thresh" refers to the threshold with which classification was performed.

Thresh	Model	Precision	Recall	F1
0.4	Baseline	0.8571	1.0	0.9231
	Proposed	**1.0**	**1.0**	**1.0**
0.5	Baseline	0.8571	1.0	0.9231
	Proposed	**1.0**	**1.0**	**1.0**
0.6	Baseline	0.9231	1.0	0.96
	Proposed	**1.0**	**1.0**	**1.0**

Experimental results shown in Table 1 indicate that incorporating textual information (i.e., two lexical features) that CNN-based models cannot read significantly enhances the performance of table detection through thresholding out FPs and alleviates dependence on the training dataset through implementing more universal features.

Acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (N0. 2021R1A2C3004345), by the Republic of Korea's MSIT (Ministry of Science and ICT), under the High-Potential Individuals Global Training Program (No. 2020-0-01463) supervised by the IITP (Institute of Information and Communications Technology Planning Evaluation), and by the Yonsei University, Republic of Korea Research Fund of 2021 (2021-22-0083).

References

[1] D. Pinto, A. McCallum, X. Wei, W. B. Croft, "Table extraction using conditional random fields", in *Proc. Annual Int. ACM SIGIR Conf. Research and Development in Inf. Retrieval (SIGIR)*, Toronto, Canada, Aug. 2003.

[2] S. Schreiber, S. Ange, I. Wolf, A. Dengel, S. Ahmed., "DeepDeSRT: Deep learning for detection and structure recognition of tables in document images," in Proc. *IAPR Int. Conf. Document Analysis and Recog. (ICDAR),* Kyoto, Japan, Nov. 2017.