TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images

Shubham Paliwal, Vishwanath D , Rohit Rahul, Monika Sharma, Lovekesh Vig
TCS Research , New Delhi
{ shubham.p3 , vishwanath.d2 , monika.sharmal , rohit.rahul , lovekesh.vig } @ tcs.com

Abstract— With the widespread use of mobile phones and scanners to photograph and upload documents, the need for extracting the information trapped in unstructured document images such as retail receipts, insurance claim forms and financial invoices is becoming more acute . A major hurdle to this objective is that these images often contain information in the form of tables and extracting data from tabular sub-images presents a unique set of challenges. This includes accurate detection of the tabular region within an image, and subsequently detecting and extracting information from the rows and columns of the detected table. While some progress has been made in table detection, extracting the table contents is still a challenge since this involves more fine grained table structure(rows & columns) recognition . Prior approaches have attempted to solve the table detection and structure recognition problems independently using two separate paper,
end deep learning model for both table detection and structure recognition. The model exploits the interdependence between the twin tasks of table detection and table structure recognition to A segment out the table and column regions. This is followed by semantic rule-based row extraction from the identified tabular sub-regions. The proposed model and extraction approach was evaluated on the publicly available ICDAR 2013 and Marmot Table datasets obtaining state of the art results . Additionally, we demonstrate that feeding additional semantic features further improves model performance and that the model exhibits transfer learning across datasets. Another contribution of this paper is to provide additional table structure annotations for the Marmot data, which currently only has annotations for table detection .

Index Terms— Table Detection, Table Structure Recognition, Scanned Documents, Information Extraction

I. INTRODUCTION

With the proliferation of mobile devices equipped with cameras, an increasing number of customers are uploading documents via these devices, making the need for information extraction from these images more pressing. Currently, these document images are often manually processed resulting in high abour costs and inefficient data processing times . Fre- quently, these documents contain data stored in tables with multiple variations in layout and visual appearance . A key component of information extraction from these documents therefore involves digitizing the data present in these tabular sub-images . The variation in the table structure, and in the graphical elements used to visually separate the tabular com- ponents make extraction from these images a very challenging problem . Most existing approaches to tabular information ex- traction divide the problem into the two separate sub-problems 1) table detection and 2) table structure recognition, and attempt to solve each sub-problem independently. While table detection involves detection of the image pixel coordinates containing the tabular sub-image, tabular structure recognition involves segmentation of the individual rows and columns in the detected table .

In this paper, we propose TableNet, a novel end-to-end deep learning model that exploits the inherent interdependence between the twin tasks of table detection and table structure identification. The model utilizes a base network that is initialized with pre-trained VGG-19 features . This is followed by two decoder branches for 1) Segmentation of the table region and 2) Segmentation of the columns within a table models. In this w region. Subsequently, rule based row extraction is employed to extract data in individual table cells .

A multi-task approach is used for the training of the deep model. The model takes a single input image and produces two different semantically labelled output images for tables and columns . The model shares the encoding layer of VGG 19 for both the table and column detectors, while the decoders for the two tasks are separate . The shared common layers are repeatedly trained from the gradients received from both the table and column detectors while the decoders are trained independently. Semantic information about elementary data types is then utilized to further boost model performance . The utilization of the VGG-19 as a base network, which is pre trained on the ImageNet dataset allows for exploitation of prior knowledge in the form of low level features learnt via training ImageNet.

We have evaluated TableNet's performance on the ICDAR 2013 dataset, demonstrating that our approach marginally outperforms other deep models as well as other state-of-the art methods in detecting and extracting tabular information in image documents . We further demonstrate that the model can generalize to other datasets with minimal fine tuning, thereby transfer learning. Furthermore, the Marmot dataset also manually annotated for column detection, and these new annotations will be publicly released to the community for future research .

In summary, the primary contributions made in this paper are as follows:

1) We propose TableNet: a novel end-to-end deep multi

task architecture for both table detection and structure recognition yielding state of the art performance on the public benchmark ICDAR and Marmot datasets .

2) We demonstrate that adding additional spatial semantic features to TableNet during training further boosts model performance .
3) We show that using a pre-trained TableNet model and fine tuning it on an another new dataset will boost the performance of the model on the new dataset, thereby allowing for transfer learning.
4) We have manually annotated the Marmot dataset for table data extraction and will release the annotations to the community

The rest of the paper is organized as follows: Section II provides an overview of the related work on tabular information extraction . Section III provides a detailed description of the TableNet model. Section IV outlines the extraction process with TableNet. Section V provides details about the datasets , preprocessing steps and training . Section VI outlines the experiment details and results . Finally, the conclusions and future work are presented in Section VII

II. RELATED WORK

There is significant prior work on identifying and extracting the tabular data inside a document . Most of these have reported results on table detection and data extraction separately (1) [2] exploited structural information to identify tables and their component fields. [3] used hierarchical representations based on the MXY tree for table detection and was the first attempt at using Machine Learning techniques for this problem . T Kasar et al. [4] identified intersecting horizontal , vertical lines and low-level features and used an SVM classifier to classify an image region as a table region or not .

Probabilistic graphical models were also used to detect tables; Silva et al. (5) modelled the joint probability distribution over sequential observations of visual page elements and the hidden state of a line (HMM) to merge potential table lines into tables resulted in a high degree of completeness. Jing Fang et al. [6] used the table header as a starting point to detect the table region and decompose its elements. Raskovic et al. [7] made an attempt to detect borderless tables. They utilized whitespaces as a heuristic rather than content for detection . After this, 19] combined deep convolutional neural networks, graphical models and saliency concepts for localizing tables and charts in documents . This technique was applied on an extended version of ICDAR 2013 table competition dataset has been very successful. and outperforms existing models . (10) locates the text components and extracts text blocks. After that , the height of each

Before the advent of deep learning, most of the work on table detection was based on heuristics or metadata . TINTIN et al. [4] identified intersecting horizontal , vertical lines and low-level features and used an SVM classifier to classify an image region as a table region or not .

Recently, DeepDeSRT (8) was proposed which uses deep learning for both table detection and table structure recognition, i.e. identifying rows, columns, and cell positions in tables can be used to detect tables and columns . The difference the detected tables . This work achieves state-of-the-art performance on the ICDAR 2013 table competition dataset .

T-Recs (11] was one of the earliest works to extract tabular based on clustering of given word segments and overlap of the text inside the table . Y. Wang et al . [12] estimates probabilities from geometric measurements made on the various entities in a given document . Ashwin et al. (13] exploit the formatting cues from semi structured HTML tables to extract data from web pages. Here the cells are already demarcated by tags since they are in HTML tables . Singh et al . [14] use object detection techniques for Document Layout understanding.

III. TABLENET: DEEP MODEL FOR TABLE AND COLUMN DETECTION

In all prior deep learning based approaches, table detection and column detection are considered separately as two different problems, which can be solved independently. However, intuitively if all the columns present in a document are table be determined known apriori, the region can easily. by definition, columns are vertically aligned word/numerical blocks. Thus , independently searching for columns can produce a lot of false positives and knowledge of the tabular region can greatly improve results for column detection (since both tables and columns have common regions). Therefore, if convolutional filters utilized to detect tables, can be reinforced by column detecting filters, this should significantly improve the performance of the model . Our proposed model , exploits this intuition and is based on the Long et al. (15) , encoder decoder model for semantic segmentation . The encoder of the model is common across both tasks, but the decoder emerges as two different branches for tables and columns . Concretely, we enforced the encoding layers to use the ground truth of both tables and columns of document for training. However , the decoding layers are separated for table and column branches. Thus, there are two computational graphs to train .

The input image for the model, is first transformed into an RGB image and then, resized to 1024 * 1024 resolution . This modified image is processed using tesseract OCR (16) as described in the previous section . Since a single model produces both the output masks for the table and column these two independent outputs have binary target pixel values, depending on whether the pixel region belongs to the table/column region or background respectively.

The problem of detecting tables in documents is similar to the problem of detecting objects in real world images . Similar to the generic object detection problem, visual features of the tolerance for noise in table/column detection is much smaller than in object detection. Therefore , instead of regressing for the boundaries of tables and columns , we employed a method to predict table and column regions pixel-wise . Recent work on semantic segmentation based on pixel wise prediction , FCN architecture, proposed by Long et al . (15) , has demonstrated the accuracy of encoder-decoder architectures for semantic segmentation . The FCN