ICDAR 2021 Scientific Literature Parsing Competition

Antonio Jimeno Yepes[1,2], Peter Zhong[1,3], and Douglas Burdick[4]

[1] IBM Research, Australia
[2] University of Melbourne, Australia
antonio.jimeno@unimelb.edu.au
[3] Oracle, Melbourne, Australia
peter.zhong@oracle.com
[4] IBM Research Almaden, USA
drburdic@us.ibm.com

Abstract. Documents in Portable Document Format (PDF) are ubiquitous with over 2.5 trillion documents. PDF format is human readable but not easily understood by machines and the large number of different styles makes it difficult to process the large variety of documents effectively.

Our ICDAR 2021 Scientific Literature Parsing Competition offers participants with a large number of training and evaluation examples compared to previous competitions. Top competition results show a significant increase in performance compared to previously reported on the competition data sets. Most of the current methods for document understanding rely on deep learning, which requires a large number of training examples. We have generated large data sets that have been used in this competition.

Our competition is split into two tasks to understand document layouts (Task A) and tables (Task B). In Task A, Document Layout Recognition, submissions with the highest performance combine object detection and specialised solutions for the different categories. In Task B, Table Recognition, top submissions rely on methods to identify table components and post-processing methods to generate the table structure and content. Results from both tasks show an impressive performance and opens the possibility for high performance practical applications.

Keywords: Document Layout Understanding · Table Recognition · ICDAR competition.

1 Introduction

Documents in Portable Document Format (PDF) are ubiquitous with over 2.5 trillion documents [8] available from several industries, including insurance documents to medical files to peer-review scientific articles. PDF represents one of the main sources of knowledge both online and offline. While PDF is great for preserving the basic elements (characters, lines, shapes, images, etc.) on a canvas for different operating systems or devices for humans to consume, it's not a format that machines can understand.

Most of the current methods for document understanding rely on deep learning, which requires a large number of training examples. We have generated large data sets automatically using PubMed Central[5] that have been used in this competition. PubMed Central is a large collection of full text articles in the biomedical domain provided by the US NIH/National Library of Medicine.

As of today, PubMed Central has almost 7 million full text articles from 2,476 journals, which offers the possibility to study the problem of document understanding over a large set of different article styles. Our data set has been generated using a subset of PubMed Central that is distributed under a Creative Commons license available for commercial use.

The competition is split in two tasks that address the understanding of document layouts by asking participants to identify several categories of information in document pages (Task A) and the understanding of tables by asking participants to produce an HTML version of table images (Task B). The IBM Research AI Leaderboard system was used to collect and evaluate the submissions of the participants. This system is based on EvalAI[6].

We had a large number of participants' submissions with over 200 for the Evaluation Phase of Task A from over 80 different teams. Results from both tasks show an impressive performance by current state-of-the-art algorithms, improving significantly over previously reported results, which opens the possibility for high performance practical applications.

2 Task A - Document Layout Recognition

This task aims to advance the research in recognizing the layout of unstructured documents. Participants of this competition need to develop a model that can identify the common layout elements in document images, including text, titles, tables, figures, and lists, with confidence score for each detection.

2.1 Related work

There has been several competitions for document layout understanding, mostly organised as ICDAR competitions. Examples of these competitions include [1], which cover as well complex layouts [3, 2], which are limited in size. There are as well data sets for document layout understanding outside competitions, for example the US NIH National Library of Medicine Medical Article Records Groundtruth (MARG) that was obtained from scanned article pages.

Overall, the previous data sets available for document layout understanding are of limited size, just several hundred pages. This is because they have been

[5] https://www.ncbi.nlm.nih.gov/pmc

[6] https://eval.ai/

manually annotation, which is slow and costly. In our Task A competition, we provide a significantly larger data set that has been generated automatically in which the validation and test sets have been manually verified.

2.2 Data

This task used the PubLayNet dataset[7] [12]. The annotations in PubLayNet are automatically generated by matching the PDF format and the XML format of the articles in the PubMed Central Open Access Subset [12].

The competition had two phases. The Format Verification Phase spanned the whole competition, for participants to verify if their results file met our submission requirements with the mini development set that we provided. The Evaluation Phase also spanned the whole competition. In this phase, participants could submit results on the test samples for evaluation. Final ranking and winning teams were decided by the performance in the Evaluation Phase. Table 2.2 shows the statistics of the data sets used in the different phases of the Task A competition.

Split	Size	Phase
Training	335,703	N/A
Development	11,245	N/A
Mini development	20	Format Verification Phase
Test	11,405	Evaluation

Table 1. Task A data set statistics

The results submitted by the participants have been objectively and quantitatively evaluated using the mean average precision (MAP) @ intersection over union (IoU) [0.50:0.95] metric on bounding boxes, which is used in the COCO object detection competition[8]. We calculated the average precision for a sequence of IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05. Then, the mean of the average precision on all element categories was computed as the final score.

2.3 Results

In the Evaluation Phase, we had more than 200 submissions from over 80 teams. Table 2 shows the top 9 results for the Evaluation Phase of the competition. Overall results and individual results are significantly higher compared to previously reported results [12]. The three top systems manage to have an overall performance above 0.97.

[7] https://github.com/ibm-aur-nlp/PubLayNet
[8] http://cocodataset.org/#detection-eval

The top performing systems, as described in the next section, relied on object detection approaches, which is similar to previous work on this data set. In addition, the predictions from object detection were compared to information extracted from the PDF version of the page or from specialized classifiers. This seems to be applied in most cases to the title and text categories, which significantly improve the performance of previously reported results.

Team Name	Text	Title	List	Table	Figure	Overall
Davar-Lab-OCR	**0.9838**	**0.9607**	0.9680	0.9735	0.9804	**0.9733**
TAL	0.9823	0.9420	**0.9700**	**0.9775**	**0.9833**	0.9710
Simo	0.9810	0.9536	0.9636	0.9738	0.9796	0.9703
BIT-VR Lab	0.9778	0.9270	0.9645	0.9762	0.9816	0.9654
IOD	0.9774	0.9251	0.9620	0.9773	0.9814	0.9647
小牛刀	0.9797	0.9515	0.9575	0.9635	0.9709	0.9646
JHL	0.9774	0.9245	0.9620	0.9754	0.9814	0.9642
刷不了	0.9778	0.9248	0.9634	0.9734	0.9803	0.9639

Table 2. Task A results

2.4 Systems description

Team: Davar-Lab-OCR, Hikvision Research Institute The system is built based on a multi-modal Mask-RCNN-based object detection framework. For a document, we make full use of the advantages from vision and semantics, where the vision is introduced in the form of document image, while semantics (texts and positions) is directly parsed from PDF. We adopt a two-stream network to extract modality-specific visual and semantic features. The visual branch processes document image and semantic branch extracts features from text embedding maps (text regions are filled with the corresponding embedding vectors, which are learned from scratch). The features are fused adaptively as the complete representation of the document, and then are fed into a standard object detection procedure.

To further improve accuracy, model ensemble technique is applied. Specifically, we train two large multimodal layout analysis models (a. ResNeXt-101-Cascade DCN Mask RCNN; b. ResNeSt-101-Cascade Mask RCNN), and inference the models in several different scales. The final results are generated by a weighted bounding-boxes fusion strategy. The code and related paper will be published in https://davar-lab.github.io/news.html.

Team: Tomorrow Advancing Life (TAL) TAL[9] used HTC (Hybrid Task Cascade for Instance Segmentation) as the baseline, which is an improved version of cascade mask rcnn. We first used some general optimization:

[9] http://www.100tal.com/about.html

(1) carefully designed the ratio of anchor;

(2) add deformable convolution module and global context block to the backbone;

(3) replace FPN with PAFPN;

(4) extract multi-level features instead of one-level features; (5) adopt IOU-balanced sampling to make the training samples more representative.

To tackle the difficulty of precise localization, we use two methods:

(1) we implement the algorithm SABL (Side-Aware Boundary Localization), where each side of the bounding box is respectively localized with a dedicated network branch;

(2) we train an expert model for the 'title' category to further improve the localization precision

In the post-processing stage, a classification model and self-developed text line detection model are used to solve the problem of missing detection in specific layout. In order to solve the problem of false detection of non target text, LayoutLM[10] is used to classify each line of text and remove the non target class.

At last, we ensemble multiple backbone models such as resnest200, resnext101, etc, and set different nms threshold for different categories. References[11] [12].

Team: Simo, Shanghai Jiao Tong University We treat the document layout analysis as an object detection task, and achieve it based on the framework of mmdetection. We first train a baseline model (Mask-RCNN). Afterwards, we improved our model from the following aspects:

1. *Annotations*: We find that for the “text” category, some samples in the train dataset are unannotated, which leads to low recall of this category. Thus we design heuristic strategies to replenish the annotations in the training dataset, which can increase the overall AP on category of “text”.

2. *Large models*: To improve performance, the network is trained based on a large backbone (ResNet-152), together with GCB and DCN blocks, which can improve our performance largely.

3. *Results refinement*: For categories of “text” and “title”, we use the coordinates extracted from the PDF to refine the final results. Specifically, we parse the text line coordinates through PDFMiner, and refine the layout prediction (large box) using the above line coordinates.

4. *Model ensemble*: Finally, we use model ensemble techniques to ensemble the above results as our final result.

Team: BIT-VR Lab In this work, our base detection method follows the two-stage framework of DetectoRS that employs HTC branch to make full use of instance and segmentation annotation to enhance the feature flow in the feature extraction stage. We train a series of CNN models based on this method with

[10] https://github.com/microsoft/unilm/tree/master/layoutlm
[11] LayoutLM: https://github.com/microsoft/unilm/tree/master/layoutlm
[12] mmdetection:https://github.com/open-mmlab

different backbones, larger input image scales, customized anchor size, various loss functions, rich data augmentation and soft-NMS method. More specifically, we use NAS technique to obtain optimal network architecture and optimal parameter configuration. Another technique is that we use OHEM to make training more effective and efficient and improve the detection accuracy of difficult samples like the "Title" category.

Besides, we trained Yolov5x model as our one-stage objection detection method, and CenteNet2 to take advantage of different characteristics in both one-stage and two-stage methods. To obtain the final ensemble detection results, we combine three different network frameworks as above and different multi-scale testing approaches with specific ensemble strategy.

3 Task B - Table Recognition

Information in tabular format is prevalent in all sorts of documents. Compared to natural language, tables provide a way to summarize large quantities of data in a more compact and structured format. Tables provide as well a format to assist readers with finding and comparing information. This competition aims to advance the research in automated recognition of tables in unstructured formats.

Participants of this task had to develop a system to convert images of tabular data into the corresponding HTML code. HTML tags that define the font style including bold, italic, strike through, superscript, and subscript had to be included in cell content. The HTML code did not need to reconstruct the appearance of tables such as border lines, background color, font, font size, or font color.

3.1 Related work

There are other table recognition challenges, which are mainly organized at the International Conference on Document Analysis and Recognition (ICDAR). ICDAR 2013 Table Competition is the first competition on table detection and recognition [5]. A total of 156 tables are included in ICDAR 2013 Table Competition for evaluation of table detection and table recognition methods; however, no training data is provided. ICDAR 2019 Competition on Table Detection and Recognition provides training, validation, and test samples (3,600 in total) for table detection and recognition [4]. Two types of documents, historical handwritten and model programmatic, are offered in image format. The ICDAR 2019 competition includes three tasks: 1) identifying table regions; 2) recognizing table structure with given table regions; 3) recognizing table structure without given table regions. The ground truth only includes the bounding box of table cell, without the cell content.

Our Task B competition proposed a more challenging task: the model needs to recognize both the table structure and the cell content of a table solely relying on the table image. In another word, the model needs to infer the tree-structure of the table and the properties (content, row-span, column-span) of each leaf node (table header/body cells). In addition, we do not provide intermediate annotations of cell position, adjacency relations, or row/column segmentation, which are needed to train most of the existing table recognition models. We only provide the final results of the tree representation for supervision. We believe this will motivate participants to develop novel models for image-to-structure mapping.

3.2 Data

This task used the PubTabNet dataset (v2.0.0)[13] [11]. PubTabNet contains over 500k training samples and 9k validation samples, of which the ground truth HTML code, and the position of non-empty table cells are provided. Participants can use the training data to train their model and the validation data for model selection and hyper-parameter tuning. The 9k+ final evaluation set (image only, no annotation) was released 3 days before the competition ended for the Final Evaluation Phase. Participants submitted their results on this set in the final phase. Submissions were evaluated using the TEDS metric[14] [11].

The competition had three phases. The Format Verification Phase spanned the whole competition, for participants to verify if their results file met our submission requirements with the mini development set that we provided. The Development Phase spanned from the beginning of the competition to 3 days before the competition ended. In this phase, participants could submit results on the test samples to verify their model. The Final Evaluation Phase run in the final 3 days of this competition. Participants could submit the inference results on the final evaluation set in this phase. Final ranking and winning teams were decided by the performance in the Final Evaluation Phase. Table 3.2 shows the size of the different data sets used in the different Task B phases.

Split	Size	Phase
Training	500,777	N/A
Development	9,115	N/A
Mini development	20	Format Verification Phase
Test	9,138	Development
Final evaluation	9,064	Final evaluation

Table 3. Task B data set statistics

3.3 Results

For Task B, we had over 30 submissions from over 30 teams for the Final Evaluation Phase. Top 10 systems ranked using their TEDS performance on the final

[13] https://github.com/ibm-aur-nlp/PubTabNet
[14] https://github.com/ibm-aur-nlp/PubTabNet/tree/master/src

evaluation set are shown in table 4. Due to a problem with the final evaluation data set, bold tags <b> where not considered in the evaluation.

The first four systems have similar performance, while we see a more significant different thereafter. As it is shown in the description of the systems, they rely on the combination of several components that identify relevant components from table images and then compose them. The performance is better than compared to previously reported result of 91 in the TEDS metric using an image to sequence approach [12].

Team Name	TEDS Simple	TEDS Complex	TEDS all
Davar-Lab-OCR	97.88	94.78	**96.36**
VCGroup	**97.90**	94.68	96.32
XM	97.60	**94.89**	96.27
YG	97.38	94.79	96.11
DBJ	97.39	93.87	95.66
TAL	97.30	93.93	95.65
PaodingAI	97.35	93.79	95.61
anyone	96.95	93.43	95.23
LTIAYN	97.18	92.40	94.84

Table 4. Task B top TEDS results. The overall result (TEDS all) is decompose into simple and complex tables [11]

3.4 Systems description

Team: Davar-Lab-OCR, Hikvision Research Institute The table recognition framework contains two main processes: table cells generation and structure inference[15].

(1) Table cells generation is built based on the Mask-RCNN detection model. Specifically, the model is trained to learn the row/column aligned cell-level bounding boxes with corresponding mask of text content region. We introduce the pyramid mask supervision and adopt a large backbone of HRNet-W48 Cascade Mask RCNN to obtain the reliable aligned bounding boxes. In addition, we train a single-line text detection model with an attention-based text recognition model to provide the OCR information. This is simply achieved by selecting the instances that only contain single-line text. We also adopt multi-scale ensemble strategy on both the cell and single-line text detection models to further improve the performance.

(2) In the structure inference stage, cells' boxes can be horizontally/vertically connected according to their alignment overlaps. The row/column information is then generated via a Maximum Clique Search process, during which the empty cells can also be located easily.

[15] Davar-Lab-OCR paper and source code: https://davar-lab.github.io.

To handle some special cases, we train another table detection model to filter out the text that do not belong to the table.

Team: VCGroup In our method [6, 7, 10][16], we divide the table content recognition task into four sub-tasks: table structure recognition, text line detection, text line recognition, and box assignment. Our table structure recognition algorithm is customized based on MASTER, a robust image text recognition algorithm. PSENet is used to detect each text line in the table image. For text line recognition, our model is also built on MASTER. Finally, in the box assignment phase, we associated the text boxes detected by PSENet with the structure item reconstructed by table structure prediction, and fill the recognized content of the text line into the corresponding item. Our proposed method achieves a 96.84% TEDS score on 9,115 validation samples in the development phase, and a 96.32% TEDS score on 9,064 samples in the Final Evaluation Phase.

Team: Tomorrow Advancing Life(TAL) The TAL system consists of two schemes:

1. Rebuild table structure through 5 detection models, which are table head-body detection, row detection, column detection, cell detection and text-row detection. Mask RCNN is selected as the baseline for these 5 detection models, with targeted optimization for different detection tasks. In the recognition part, the results of cell detection and text-row detection are inputted into the CRNN model to get the recognition result corresponding to each cell.

2. The restoration of table structure is treated as an img2seq problem. To shorten the decoding length, we replace every cell content with different numbers. The numbers are obtained from text-row detection results. Then we use CNN to encode the image and use a transformer model to decode the structure of the table. The corresponding text-line content can then be obtained by using the CRNN model.

The above two schemes can be used to get the complete table structure and content recognition results. We have a set of selection rules, which combine the advantages of both schemes, to output the one best final result.

Team: PaodingAI, Beijing Paoding Technology Co., Ltd PaodingAI's system is divided into three main parts: text block detection, text block recognition, and table structure recognition. The text block detector is trained by the Detectors_cascade_rcnn_r50_2x model provided by MMDetection. The text block recognizer is trained by the SAR_TF [17] model. Table structure recognizer is our own implementation of the model proposed in [9]. In addition to the above model, we also use rules and a simple classification model to process <thead>, <b>, and blank characters. Our system is not an end-to-end model and does not use an integrated approach.

[16] VCGroup Github repo: https://github.com/wenwenyu/MASTER-pytorch

[17] https://github.com/Pay20Y/SAR_TF

4 Conclusions

We have proposed two tasks for document understanding using large data sets derived from PubMed Central for the training and evaluation of participant systems. These tasks address two important problems, understanding document layouts and tables.

We had a large participation for both tasks, which was quite significant for Task A with over 200 submissions from over 80 teams. Results from top participant submissions significantly improve the performance of previously reported results. Results from both tasks show an impressive performance and opens the possibility for high performance practical applications.

5 Acknowledgements

We would like to thank Sundar Saranathan for his help with the competition system.

We would like to thank the US NIH/National Library of Medicine for making available the data sets used in this competition.

References

1. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 296–300. IEEE (2009)
2. Clausner, C., Antonacopoulos, A., Pletschacher, S.: Icdar2017 competition on recognition of documents with complex layouts-rdcl2017. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1404–1410. IEEE (2017)
3. Clausner, C., Papadopoulos, C., Pletschacher, S., Antonacopoulos, A.: The enp image and ground truth dataset of historical newspapers. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 931–935. IEEE (2015)
4. Gao, L., Huang, Y., Li, Y., Yan, Q., Fang, Y., Dejean, H., Kleber, F., Lang, E.M.: ICDAR 2019 competition on table detection and recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1510–1515. IEEE (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00166
5. Göbel, M., Hassan, T., Oro, E., Orsi, G.: ICDAR 2013 table competition. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1449–1453. IEEE (2013)
6. He, Y., Qi, X., Ye, J., Gao, P., Chen, Y., Li, B., Tang, X., Xiao, R.: Pingan-vcgroup's solution for icdar 2021 competition on scientific table image recognition to latex. arXiv (2021)
7. Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: Master: Multi-aspect non-local network for scene text recognition. Pattern Recognition (2021)

8. Staar, P.W., Dolfi, M., Auer, C., Bekas, C.: Corpus conversion service: A machine learning platform to ingest documents at scale. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 774–782 (2018)
9. Tensmeyer, C., Morariu, V.I., Price, B., Cohen, S., Martinez, T.: Deep splitting and merging for table structure decomposition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 114–121. IEEE (2019)
10. Ye, J., Qi, X., He, Y., Chen, Y., Gu, D., Gao, P., Xiao, R.: Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: Table recognition to html. arXiv (2021)
11. Zhong, X., ShafieiBavani, E., Yepes, A.J.: Image-based table recognition: data, model, and evaluation. arXiv preprint arXiv:1911.10683 (2019)
12. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1015–1022. IEEE (2019)