DOCUMENT AI: BENCHMARKS, MODELS AND APPLICATIONS

Lei Cui, Yiheng Xu, Tengchao Lv, Furu Wei
Microsoft Research Asia
{lecu,t-yihengxu,tengchaolv,fuwei}@microsoft.com

ABSTRACT

Document AI, or Document Intelligence, is a relatively new research topic that refers to the techniques for automatically reading, understanding, and analyzing business documents. It is an important research direction for natural language processing and computer vision. In recent years, the popularity of deep learning technology has greatly advanced the development of Document AI, such as document layout analysis, visual information extraction, document visual question answering, document image classification, etc. This paper briefly reviews some of the representative models, tasks, and benchmark datasets. Furthermore, we also introduce early-stage heuristic rule-based document analysis, statistical machine learning algorithms, and deep learning approaches especially pre-training methods. Finally, we look into future directions for Document AI research.

1 DOCUMENT AI

Document AI, or Document Intelligence, is a booming research topic with increased industrial demand in recent years. It mainly refers to the process of automated understanding, classifying and extracting information with rich typesetting formats from webpages, digital-born documents or scanned documents through AI technology. Due to the diversity of layouts and formats, low-quality scanned document images, and the complexity of the template structure, Document AI is a very challenging task and has attracted widespread attention in related research areas. With the acceleration of digitization, the structured analysis and content extraction of documents, images and others has become a key part of the success of digital transformation. Meanwhile automatic, accurate, and rapid information processing is crucial to improving productivity. Taking business documents as an example, they not only contain the processing details and knowledge accumulation of a company's internal and external affairs, but also a large number of industry-related entities and digital information. Manually extracting information is time-consuming and labor-intensive with low accuracy and low reusability. Document AI deeply combines artificial intelligence and human intelligence, and has different types of applications in multiple industries such as finance, healthcare, insurance, energy and logistics. For instance, in the finance field, it can conduct financial report analysis and intelligent decision analysis, and provide scientific and systematic data support for the formulation of corporate strategies and investment decisions. In healthcare, it can improve the digitization of medical cases and enhance diagnosis accuracy. By analyzing the correlation between medical literature and cases, people can locate potential treatment options. In the accounting field, it can achieve automatic information extraction of invoices and purchase orders, automatically analyze a large number of unstructured documents, and support different downstream business scenarios, saving a lot of manual processing time.

Over the past few decades, the development of document intelligence has roughly gone through different stages, evolving from simple rule-based heuristics to neural network approaches. In the early 1990s, researchers mostly used rule-based heuristic approaches for document understanding and analysis. By manually observing the layout information of documents, they summarized some heuristic rules and processed documents with fixed layout information. However, traditional rule-based methods often require large labor costs, and these manually summarized rules are not scalable. Therefore, researchers have begun to adopt methods based on statistical machine learning. Since 2000, with the development of machine learning technology, machine learning models based on annotated data have become the mainstream method of document processing. It uses artificially
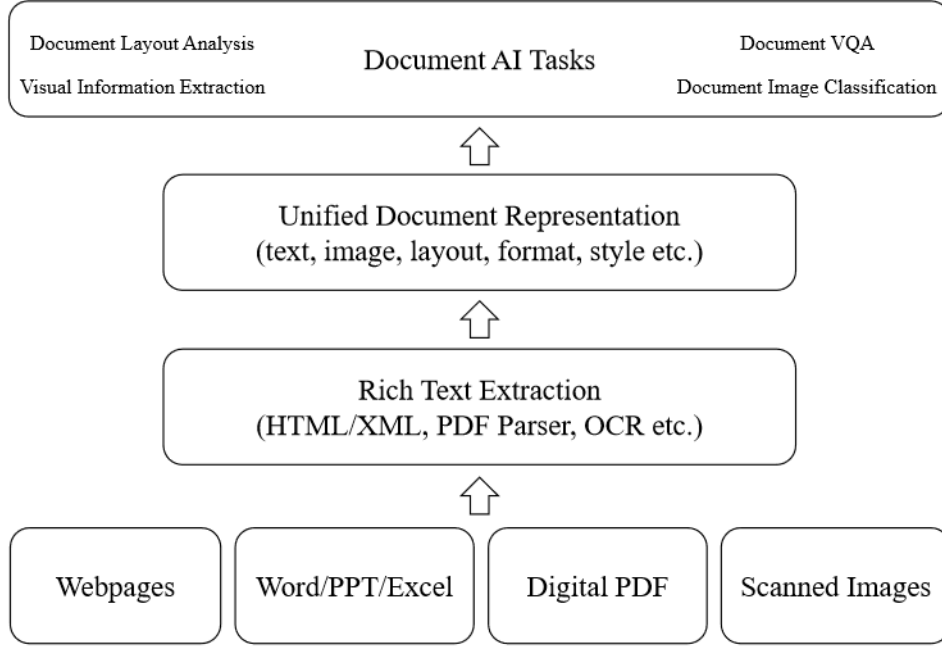
Figure 1: Overview of Document AI

designed feature templates to learn the weights of different features to understand and analyze the content and layout of a document. Although annotated data is leveraged for supervised learning and previous methods can bring a certain degree of performance improvement, the general usability is often not satisfactory due to the lack of customized rules and the number of training samples. Additionally, the migration and adaptation costs for different types of documents are relatively high, making previous approaches not practical for widespread commercial use. In recent years, with the development of deep learning technology and the accumulation of a large number of unlabeled electronic documents, document analysis and recognition technology has entered a new era. Figure 1 represents the basic framework of Document AI technology under the current deep learning framework, in which different types of documents are extracted through content extraction tools (HTML/XML extraction, PDF parser, OCR, etc.) where text content, layout information, and visual image information are well organized. Then, large-scale deep neural networks are pre-trained and fine-tuned to complete a variety of downstream Document AI tasks, including document layout analysis, visual information extraction, document visual question answering, and document image classification etc. The emergence of deep learning, especially of the pre-training technique represented by Convolutional Neural Networks (CNN), Graph Neural Networks (GNN) and Transformer architecture (Vaswani et al., 2017), has completely shifted the traditional machine learning paradigm that requires a lot of manual annotations. Instead, it heavily relies on a large amount of unlabeled data for self-supervised learning, and addresses the downstream tasks through the "pre-training and fine-tuning" paradigm which leads to a significant breakthrough in Document AI tasks. We have also observed many successful Document AI products, such as Microsoft Form Recognizer[1], Amazon Textract[2], Google Document AI[3] and many others, which have fundamentally empowered a variety of industries with the Document AI technology.

Although deep learning has greatly improved the accuracy of Document AI tasks, there are still many problems to be solved in practical applications. First, due to the limitation of the input length of current large-scale pre-trained language models, they usually need to truncate documents into several parts to be input to the model for processing, which poses a great challenge for the multi-page and cross-page understanding of complex long documents. Second, due to the quality mismatch between annotated training data and document images in real-world business which usually comes

---

[1]https://azure.microsoft.com/en-us/services/form-recognizer/
[2]https://aws.amazon.com/textract
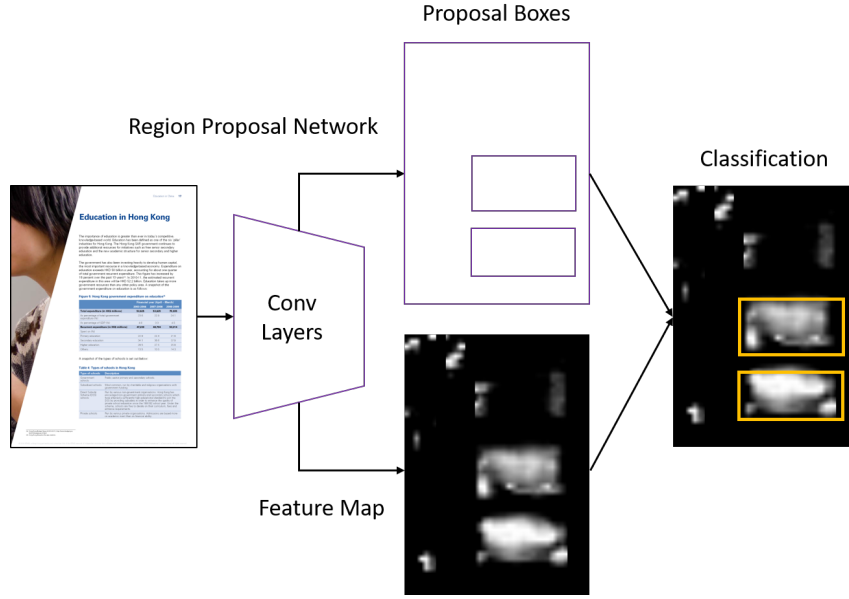[3]https://cloud.google.com/document-ai

Figure 2: Document layout analysis with Faster R-CNN

from the scanning equipment, crumpled paper and random placement, poor performance is observed and more data synthesis/augmentation techniques are needed to help existing models improve the performance. Third, the current Document AI tasks are often trained independently, and the correlation between different tasks has not been effectively leveraged. For instance, visual information extraction and document visual question answering have some common semantic representations, which can be better solved by using a multi-task learning framework. Finally, the pre-trained Document AI models also encountered the problem of insufficient computing resources and labeled training samples in practical applications. Therefore, model compression, few-shot learning and zero-shot learning are important research directions at present and have great practical values.

Next, we introduce the current mainstream Document AI models (CNN, GNN and Transformer), tasks, and benchmark datasets, and then elaborate on early-stage document analysis techniques based on heuristic rules, algorithms, and models based on traditional statistical machine learning, as well as the most recent deep learning models, especially the multimodal pre-training technique. Finally, we outline future directions of Document AI research.

2 REPRESENTATIVE MODELS, TASKS AND BENCHMARKS

2.1 DOCUMENT LAYOUT ANALYSIS WITH CONVOLUTIONAL NEURAL NETWORKS

In recent years, convolutional neural networks have achieved great success in the field of computer vision, especially the supervised pre-training model ResNet (He et al., 2015) based on large-scale annotated datasets ImageNet and COCO has brought great performance improvements in image classification, object detection and scene segmentation. Specifically, with multi-stage models such as Faster R-CNN (Ren et al., 2016) and Mask R-CNN (He et al., 2018), and single-stage detection models including SSD (Liu et al., 2016) and YOLO (Redmon & Farhadi, 2018), object detection has almost become a solved problem in computer vision. Document layout analysis can essentially be regarded as an object detection task for document images. Basic units such as headings, paragraphs, tables, figures and charts in the document are the objects that need to be detected and recognized. Yang et al. (2017a) regard document layout analysis as a pixel-level segmentation task, and used convolutional neural networks for pixel classification to achieve good results. Schreiber et al. (2017) first apply the Faster R-CNN model to table detection and recognition in document layout analysis as shown in Figure 2, achieving SOTA performance in the ICDAR 2013 table detection dataset (Göbel et al., 2013). Although document layout analysis is a classic document intelligence task, it has been
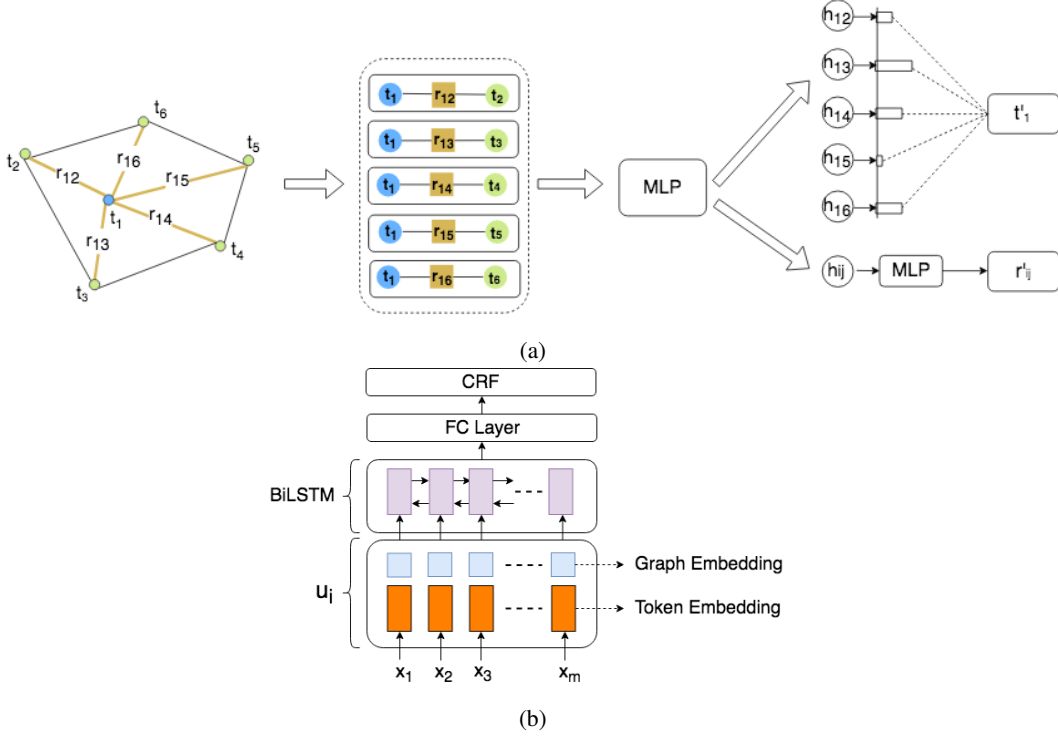
Figure 3: Visual information extraction with GNN

limited to a small training dataset for many years, which is not sufficient for applying pre-trained models in computer vision. With large-scale weakly supervised document layout analysis datasets such as PubLayNet (Zhong et al., 2019b), PubTabNet (Zhong et al., 2019a), TableBank (Li et al., 2020a) and DocBank (Li et al., 2020b), researchers can conduct a more in-depth comparison and analysis of different computer vision models and algorithms, and further promote the development of document layout analysis techniques.

2.2 VISUAL INFORMATION EXTRACTION WITH GRAPH NEURAL NETWORKS

Information extraction is the process of extracting structured information from unstructured text, and it has been widely studied as a classical and fundamental NLP problem. Traditional information extraction focuses on extracting entity and relationship information from plain text, but less research has been done on visually rich documents. Visually rich documents refers to the text data whose semantic structure is not only determined by the content of the text, but also related to visual elements such as layout, typesetting formats as well as table/figure structures. Visually-rich documents can be found everywhere in real-world applications, such as receipts, certificates, insurance files, etc. Liu et al. (2019a) propose modeling visually rich documents using graph convolutional neural networks. As shown in Figure 3, each image is passed through the OCR system to obtain a set of text blocks, each of which contains information about its coordinates in the image with the text content. This work constitutes this set of text blocks as a fully connected directed graph, i.e., each text block constitutes a node, and each node is connected to all other nodes. The initial features of the nodes are obtained from the text content of the text blocks by Bi-LSTM encoding. The initial features of the edges are the relative distance between the neighboring text blocks and the current text block and the aspect ratio of these two text blocks. Unlike other graph convolution models that only convolve on nodes, this work focuses more on the "individual-relationship-individual" ternary feature set in information extraction, so convolution is performed on the "node-edge-node" ternary feature set. In addition, the self-attention mechanism allows the network to select more noteworthy information in all directed triads in fully connected directed graphs and aggregate the weighted features. The initial node features and edge features are convolved in multiple layers to obtain the high-level representations of nodes and edges. Experiments show that this graph convolution model
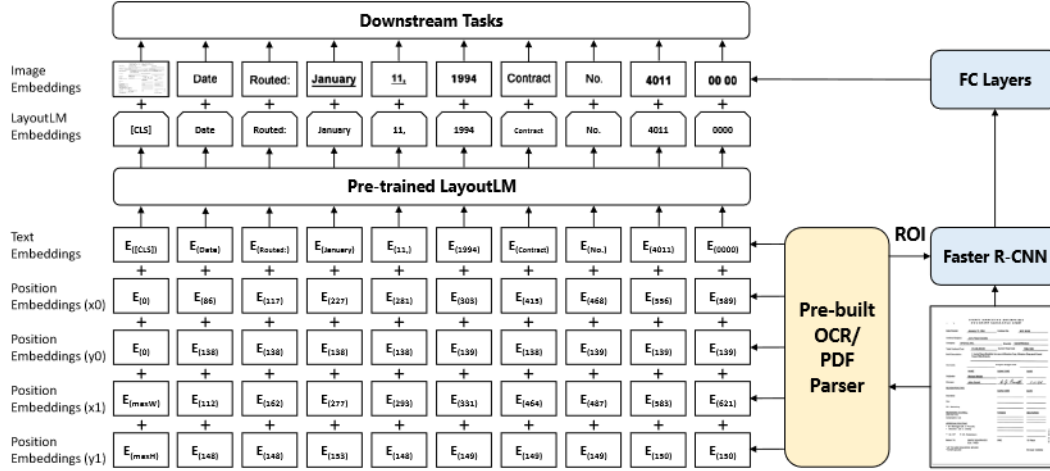
4

Figure 4: The LayoutLM architecture, where 2-D layout and image embeddings are integrated into the Transformer architecture.

significantly outperforms the Bi-LSTM+CRF models. In addition, experiments have shown that visual information plays a major role, increasing the discrimination of texts with similar semantics. Text information also plays a certain auxiliary role to visual information. The self-attention mechanism is basically not helpful for fixed layout data, but it generates some level of improvement on non-fixed layout data.

2.3 GENERAL-PURPOSE MULTIMODAL PRE-TRAINING WITH THE TRANSFORMER ARCHITECTURE

In many cases, the spatial relationship of text blocks in a document usually contains rich semantic information. For instance, forms are usually displayed in the form of key-value pairs. Typically, the arrangement of key-value pairs is usually in the left-right order or the up-down order. Similarly, in a tabular document, the text blocks are usually arranged in a grid layout and the header usually appears in the first column or row. This layout invariance among different document types is a critical property for general-purpose pre-training. Through pre-training, the position information that is naturally aligned with the text can provide richer semantic information for downstream tasks. For visually-rich documents, in addition to positional information, the visual information presented with the text can also help downstream tasks, such as font types, sizes, styles and other visually-rich formats. For instance, in forms, the key part of a key-value pair is usually given in bold form. In general documents, the title of the article will usually be enlarged and bold, and the nouns of special concepts will be displayed in italics, etc. For document-level tasks, the overall visual signals can provide global structural information, and there is a clear visual difference between different document types, such as a personal resume and a scientific paper. The visual features displayed in these visually-rich documents can be extracted by visual encoders and combined into the pre-training stage, thereby effectively improving downstream tasks.

To leverage the layout and visual information, Xu et al. (2020) propose a general document pre-training model LayoutLM (Xu et al., 2020), as shown in Figure 4. Two new embedding layers, 2-D position embedding and image embedding are added on the basis of the existing pre-trained model, so that the document structure and visual information can be effectively combined. Specifically, according to the text bounding boxes obtained by OCR, the algorithm first gets the coordinates of the text in the document. After converting the corresponding coordinates into virtual coordinates, the model calculates the representation of the coordinates corresponding to the four embedding sub-layers of x, y, w, and h. The final 2-D position embedding is the sum of the embedding of the four sub-layers. In image embedding, the model considers the bounding boxes corresponding to each text as the proposal in the Faster R-CNN to extract the corresponding local features. In particular, since the [CLS] symbol is used to represent the semantics of the entire document, the model also uses the entire document image as the image embedding at this position to maintain multimodal alignment.

In the pre-training stage, the authors propose two self-supervised pre-training tasks for LayoutLM:

Task #1: Masked Visual-Language Model   Inspired by the masked language model, the authors propose the Masked Visual-language Model (MVLM) to learn language representation with the clues of 2-D position embeddings and text embeddings. During pre-training, the model randomly masks some of the input tokens but keeps the 2-D position embeddings and other text embeddings. The model is then trained to predict the masked tokens given the context. In this way, the LayoutLM model not only understands the language contexts, but also utilizes the corresponding 2-D position information, thereby bridging the gap between the visual and language modalities.

Task #2: Multi-Label Document Classification   For document image understanding, many tasks require the model to generate high-quality document-level representations. As the IIT-CDIP Test Collection includes multiple tags for each document image, the model also uses a Multi-label Document Classification (MDC) loss during the pre-training phase. Given a set of scanned documents, the model uses the document tags to supervise the pre-training process so that the model can cluster the knowledge from different domains and generate better document-level representation. Since the MDC loss needs the label for each document image that may not exist for larger datasets, it is optional during the pre-training and may not be used for pre-training larger models in the future.

Experiments show that pre-training with layout and visual information can be effectively transferred to downstream tasks. Significant accuracy improvements are achieved in multiple downstream tasks. Different from the convolutional neural networks and graph neural networks, the advantage of the general document-level pre-training model is that it can support different types of downstream applications.

2.4 Mainstream Document AI Tasks and Benchmarks

Document AI involves automatic reading, comprehension, and analysis of documents. In real-world application scenarios, it mainly includes four types of tasks, namely:

Document Layout Analysis   This is the process of automatic analysis, recognition and understanding of images, text, table/figure/chart information and positional relationships in the document layout.

Visual Information Extraction   This refers to the techniques of extracting entities and their relationships from a large amount of unstructured content in a document. Unlike traditional pure text information extraction, the construction of the document turns the text from a one-dimensional sequential arrangement into a two-dimensional spatial arrangement. This makes text information, visual information and layout information extremely important influencing factors in visual information extraction.

Document Visual Question Answering   Given digital-born documents or scanned images, PDF parsing, OCR or other text extraction tools are first used to automatically recognize the textual content, the system needs to answer the natural language questions about the documents by judging the internal logic of the recognized text.

Document Image Classification   This refers to the process of analyzing and identifying document images, while classifying them into different categories such as scientific papers, resumes, invoices, receipts and many others.

For these four main Document AI tasks, there have been a large number of open-sourced benchmark datasets in academia and industry, as shown in the Table 1. This has greatly promoted the construction of new algorithms and models by researchers in related research areas, especially the most recent deep learning based models that achieve SOTA performance in these tasks. Next, we will introduce in detail the classic models and algorithms in different periods in the past, including document analysis techniques based on heuristic rules, document analysis technology based on statistical machine learning, and general Document AI models based on deep learning.

Task	Benchmark	Langauge	Paper/Link
Document Layout Analysis	ICDAR 2013	En	Göbel et al. (2013)
	ICDAR 2019	En	Gao et al. (2019)
	ICDAR 2021	En	Yepes et al. (2021)
	UNLV	En	Shahab et al. (2010)
	Marmot	Zh/En	Fang et al. (2012)
	PubTabNet	En	Zhong et al. (2019a)
	PubLayNet	En	Zhong et al. (2019b)
	TableBank	En	Li et al. (2020a)
	DocBank	En	Li et al. (2020b)
	TNCR	En	Abdallah et al. (2021)
	TabLeX	En	Desai et al. (2021)
	PubTables	En	Smock et al. (2021)
	IIIT-AR-13K	En	Mondal et al. (2020)
	ReadingBank	En	Wang et al. (2021b)
Visual Information Extraction	SWDE	En	Hao et al. (2011)
	FUNSD	En	Guillaume Jaume (2019)
	SROIE	En	Huang et al. (2019)
	CORD	En	Park et al. (2019)
	EATEN	Zh	Guo et al. (2019)
	EPHOIE	Zh	Wang et al. (2021a)
	Deepform	En	Stray & Svetlichnaya (2020)
	Kleister	En	Stanisławek et al. (2021)
	XFUND	Zh/Ja/Es/Fr/It/De/Pt	Xu et al. (2021b)
Document VQA	DocVQA	En	Mathew et al. (2021b)
	InfographicsVQA	En	Mathew et al. (2021a)
	VisualMRC	En	Tanaka et al. (2021)
	WebSRC	En	Chen et al. (2021)
	Insurance VQA	Zh	`https://bit.ly/36O2Vow`
Document Image Classification	Tobacco-3482	En	Kumar et al. (2014)
	RVL-CDIP	En	Harley et al. (2015)

Table 1: Benchmark datasets for document layout analysis, visual information extraction, document visual question answering and document image classification.

3 Heuristic Rule-based Document Layout Analysis

Document layout analysis using heuristic rules can be roughly divided into three ways: top-down, bottom-up, and hybrid strategy. The top-down methods divide a document image into different areas step by step. Cutting is performed recursively until the area is divided to a predefined standard, usually blocks or columns. The bottom-up methods use pixels or components as the basic element units, where the basic elements are grouped and merged to form a larger homogeneous area. The top-down approach enables faster and more efficient analysis of documents in specific formats, while bottom-up approaches require more computation resources but are more versatile and can cover more documents with different layout types. The hybrid strategy combines the top-down and bottom-up to produce better results.

This section introduces document analysis techniques from the top-down and bottom-up perspectives, including projection profile, image smearing, connected components and others.

3.1 Projection Profile

ccProjection profile is widely used in document analysis as a top-down analysis method. Nagy & Seth (1984) use the X-Y cut algorithm to cut the document. This method is suitable for structured text with fixed text areas and line spacing, but it is sensitive to boundary noise and cannot provide good results on slanted text. Bar-Yosef et al. (2009) use the dynamic local projection-profile to calculate the inclination of the document in an attempt to eliminate performance degradation caused

by text skew. Experiments have proven that the model has obtained more accurate results on slanted and curved text. In addition, many variations of the X-Y cut algorithm have been proposed to address existing problems in document analysis. O'Gorman (1993) extends the X-Y cut algorithm to use the projection of the component bounding boxes, and Sylwester & Seth (1995) use an evaluation metric called edit-cost to guide the segmentation model, which improves overall performance.

The projection profile analysis is suitable for structured text, especially documents with a manhattan-based layout. The performance may not be satisfactory for documents with complex layouts, slanted text, or border noises.

3.2 Image Smearing

Image smearing refers to permeating from one location to the surroundings and gradually expanding to all homogeneous areas to determine its layout in the page. Wong et al. (1982) adopt a top-down strategy and uses Run-Length Smoothing Algorithm (RLSA) to determine homogeneous regions. After the image is binarized, the pixel value 0 represents the background, and 1 is the foreground. When the number of 0s around 0 is less than the specified threshold C, the 0 at that position is changed to 1, and the RLSA uses this operation to merge the foreground that is nearby into a whole unit. In this way, characters can be gradually merged into words, and words can be merged into lines of text, and then the range continues to extend to the entire homogeneous area. On this basis, Fisher et al. (1990) go further by adding preprocessing such as noise removal and tilt correction. In addition, the threshold C of RLSA is modified according to the dynamic algorithm to further improve adaptability. Esposito et al. (1990) use a similar approach but the operation object is changed from pixels to character frames. Shi & Govindaraju (2004) expand each position pixel in the image to obtain a new grayscale image, which is extracted and shows good performance in the case of handwritten fonts, text slanting, etc.

3.3 Connected Components

As a bottom-up approach, connected components infers the relationship among the small elements, which is used to find homogeneous regions and finally classifies the regions into different layouts. Fisher et al. (1990) use connected components to find the K-Nearest Neighbors (KNN) components of each component and infer the attributes of the current area through the relationship between the positions and angles of each other. Saitoh et al. (1993) merge the text into lines according to the inclination of the document, and then merge the lines into regions and classify them into different attributes. Kise et al. (1998) also try to solve the problem of text skew. The authors use an approximated area Voronoi diagram to obtain the candidate boundary of the area. This operation is effective for areas with any angle of inclination. However, due to the need to estimate character spacing and line spacing during the calculation process, the model cannot perform well when the document contains large fonts and wide character spacing. In addition, Bukhari et al. (2010) also use AutoMLP on the basis of connected components in order to find the best parameters of the classifier to further improve the performance.

3.4 Other Approaches

In addition to the above methods, there are some other heuristic rule based document layout analysis approaches. Baird et al. (1990) use a top-down approach to divide the document into areas by blanks. Xiao & Yan (2003) use the Delaunay Triangulation algorithm for document analysis. On this basis, Bukhari et al. (2009) apply it to script-independent handwritten documents. In addition, there are some hybrid models. Okamoto & Takahashi (1993) use separators and blanks to cut blocks, and further merge internal components into text lines in each block. Smith (2009) divide the document analysis into two parts. First, the bottom-up method is used to locate the tab characters, and the column layout is inferred with the help of the tab characters. Second, it uses a top-down approach to infer the structure and text order on the column layout.

4 MACHINE LEARNING BASED DOCUMENT LAYOUT ANALYSIS

The machine learning based document analysis process is usually divided into two stages: 1) segmenting the document image to obtain multiple candidate regions; 2) classifying the document regions and distinguishing them such as text blocks and images. Some research work tries to use machine learning algorithms for document segmentation, while the rest try to construct features on generated regions and use machine learning algorithms to classify the regions. In addition, due to the performance boost led by machine learning, more machine learning models have been tried in table detection tasks, since table detection is a vital subtask of document analysis. This section will introduce machine learning approaches for different layout analysis tasks.

4.1 DOCUMENT SEGMENTATION

For document segmentation, Baechler & Ingold (2011) combine the X-Y cut algorithm and use logistic regression to segment the document and discard the blank areas. After obtaining the corresponding regions, they also compare the performance of algorithms such as KNN, logistic regression and Maximum Entropy Markov Models (MEMM) as classifiers. The experiment shows that MEMM and logistic regression have better performance on classification tasks. Esposito et al. (2008) further strengthen machine learning algorithms in document segmentation. In a bottom-up way, a kernel-based algorithm (Dietterich et al., 1997) is used in the process of merging letters to words and text lines, and the results are converted into an xml structure for storage. After that, the Document Organization Composer (DOC) algorithm is used to analyze the documents. Wu et al. (2008) focus on the problem of two reading orders of text at the same time. The existing models assume that the text information has only one reading order, but it cannot work normally when it encounters texts written in horizontal or vertical directions, such as in Chinese or Japanese. The proposed model divides the document segmentation process into four steps for judging and processing the text, and used the Support Vector Machines (SVM) model to decide whether to execute these steps in a pre-defined order.

4.2 REGION CLASSIFICATION

For region classification, conventional research work usually leverages machine learning models to distinguish different regions. Wei et al. (2013) compare the advantages and disadvantages of SVM, multi-layer perceptrons (MLP) and Gaussian Mixture Models (GMM) as classifiers. Experiments show that the classification accuracy of SVM and MLP are significantly better than GMM in region classification. Bukhari et al. (2012) manually construct and extract multiple features from the document regions, and then used the AutoMLP algorithm to classify them. The classification accuracy of 95% is obtained in the Arabic dataset. Baechler & Ingold (2011) further improve the performance in region classification using a pyramid algorithm by conducting three levels of analysis on medieval manuscripts and using Dynamic Multi-Layer Perceptron (DMLP) as the classification model.

4.3 TABLE DETECTION

In addition to the above methods, there is a lot of research using traditional machine learning models for table detection and recognition. (Wang et al., 2000; Wangt et al., 2001; Wang et al., 2002) use a binary tree to analyze the document in a top-down way to find the candidate table areas, and determine the final table area according to the predefined features. Pinto et al. (2003) use a Conditional Random Field (CRF) model to extract the table area in the HTML page, and identify the title, subtitle and other content in the table. e Silva (2009) uses Hidden Markov Models (HMM) to extract table regions. Chen & Lopresti (2011) retrieve the table area in the handwritten document and use an SVM model to identify the texts within that region, and predict the location of the table based on the text lines. Kasar et al. (2013) identify the horizontal and vertical lines in the figure, and then use an SVM model to classify the attributes of each line to determine whether the line belongs to the table. Barlas et al. (2014) use an MLP model to classify the connected component in the document and determined whether it belongs to tables. Bansal et al. (2014) use the leptonica library Bloomberg (1991) to segment the document, and then construct features containing surrounding information for each region. By using the fix-point model Li et al. (2013) to identify the table areas, the model does not only conduct the region classification, but also learns the relationship among different areas.

Rashid et al. (2017) take the region classification into the word level and then used AutoMLP to determine whether the word belongs to the table.

5 Deep Learning based Document AI

In recent years, deep learning methods have become a new paradigm for solving many machine learning problems. Deep learning methods have been confirmed to be effective in many research areas. Recently, the popularity of pre-trained models has further improved the performance of deep neural networks. The development of Document AI also reflects a similar trend with other applications in deep learning. In this section, we divide the existing models into two parts: deep learning models for specific tasks and general-purpose pre-trained models that support a variety of downstream tasks.

5.1 Task-specific Deep Learning Models

5.1.1 Document Layout Analysis

Document layout analysis includes two main subtasks: visual analysis and semantic analysis (Binmakhashen & Mahmoud, 2019). The main purpose of visual analysis is to detect the structure of the document and determine the boundaries of similar regions. Semantic analysis needs to identify specific document elements, such as headings, paragraphs, tables, etc., for these detected areas. PubLayNet (Zhong et al., 2019b) is a large-scale document layout analysis dataset. More than 360,000 document images are constructed by automatically parsing PubMed XML files. DocBank (Li et al., 2020b) automatically built an extensible document layout analysis dataset through the correspondence between PDF files and LaTeX files on the arXiv website, and supports both text-based and image-based document layout analysis. IIIT-AR-13K (Mondal et al., 2020) also provided 13,000 manually annotated document images for layout analysis.

In Section 2.1, we introduced the application of the Convolutional Neural Network (CNN) in document layout analysis (He et al., 2015; Ren et al., 2016; He et al., 2018; Liu et al., 2016; Redmon & Farhadi, 2018; Yang et al., 2017a; Schreiber et al., 2017). As the performance requirement for document layout analysis has gradually increased, more research work has made significant improvements with specific detection models. Yang et al. (2017b) treat the document semantic structure analysis task as a pixel-by-pixel classification problem. They propose a multimodal neural network that considers both visual and textual information. Viana & Oliveira (2017) propose a lightweight model for document layout analysis of mobile and cloud services. This model uses the one-dimensional information of the image for inference, and achieves higher accuracy compared with the model that uses the two-dimensional information. Chen et al. (2017) introduce a page segmentation method of handwritten historical document images based on CNN. Oliveira et al. (2018) propose a multi-task pixel-by-pixel prediction model based on CNN. Wick & Puppe (2018) propose a high-performance Fully Convolutional Neural Network (FCN) for historical document segmentation. Grüning et al. (2019) propose a two-stage text line detection method for historical documents. Soto & Yoo (2019) incorporate contextual information into the Faster R-CNN model. This model used the local invariance of the article elements to improve the region detection performance.

Table Detection and Recognition In document layout analysis, table understanding is an important and challenging subtask. Different from document elements such as headings and paragraphs, the format of the table is usually more variable and the structure is more complex. Therefore, there is a lot of related work carried out around tables, among which the two most important subtasks are table detection and table structure recognition. 1) Table detection refers to determine the boundary of the tables in the document. (2) Table structure recognition refers to extracting the semantic structure of the table, including information about rows, columns, and cells, according to a predefined format.

In recent years, benchmark datasets have emerged for table understanding, including table detection datasets such as Marmot (Fang et al., 2012) and UNLV (Shahab et al., 2010). Meanwhile, the ICDAR conference held several competitions on table detection and recognition, where high-quality table datasets are provided (Göbel et al., 2013; Gao et al., 2019). However, these traditional benchmark datasets are relatively small in scale, and it is difficult to unleash the capability of the deep

neural networks. Therefore, TableBank (Li et al., 2020a) uses LaTex and Office Word documents to automatically build a large-scale table understanding dataset. PubTabNet (Zhong et al., 2019a) proposes a large-scale table dataset and provides table structure and cell content to assist in table recognition. TNCR (Abdallah et al., 2021) provides the label of the table categories while providing the table boundaries.

Many deep learning based object detection models have achieved good results in table detection. Faster R-CNN (Ren et al., 2016) achieves very good performance by directly applying it to table detection. On this basis, Siddiqui et al. (2018) achieve better performance by applying deformable convolution on Faster R-CNN. CascadeTabNet (Prasad et al., 2020) uses the Cascade R-CNN (Cai & Vasconcelos, 2018) model to perform table detection and table structure recognition at the same time. TableSense (Dong et al., 2019) significantly improves table detection capabilities by adding cell features and sampling algorithms.

In addition to the above two main subtasks, the understanding of parsed tables has become a new challenge. TAPAS (Herzig et al., 2020) introduces pre-training techniques to table comprehension tasks. By introducing an additional positional encoding layer, TAPAS enables the Transformer (Vaswani et al., 2017) encoder to accept structured table input. After pre-training on a large amount of tabular data, TAPAS significantly surpasses traditional methods in a variety of downstream semantic analysis tasks for tables. Following TAPAS, TUTA (Wang et al., 2020a) introduces a two-dimensional coordinate to represent the hierarchical information of a structured table, and proposes a tree structure based location representation and attention mechanism to show the hierarchical modeling of this structure. Combining different levels of pre-training tasks, TUTA has achieved further performance improvements on multiple downstream datasets.

5.1.2 VISUAL INFORMATION EXTRACTION

Visual information extraction refers to the technology of extracting semantic entities and their relationships from a large number of unstructured visually-rich documents. Visual information extraction differs in different document categories and the extracted entities are also different. FUNSD (Guillaume Jaume, 2019) is a form understanding dataset that contains 199 forms, where each sample contains key-value pairs of form entities. SROIE (Huang et al., 2019) is an OCR and information extraction benchmark for receipt understanding, which attracts a lot of attention from the research/industry community. CORD (Park et al., 2019) is a receipt understanding dataset that contains 8 categories and 54 subcategories of entities. Kleister (Stanisławek et al., 2021) is a document understanding dataset for long and complex document entity extraction tasks, including long text documents such as agreements and financial statements. DeepForm (Stray & Svetlichnaya, 2020) is an English dataset for the disclosure form of political advertisements on television. The EATEN dataset (Guo et al., 2019) is a dataset for information extraction of Chinese documents. Yu et al. (2021) further add text box annotations to the 400 subset of EATEN. The EPHOIE (Wang et al., 2021a) dataset is also an information extraction dataset for Chinese document data. XFUND (Xu et al., 2021b) is a multi-lingual extended version of the FUNSD data set proposed with the LayoutXLM model, which contains visually-rich documents in seven commonly-used languages.

For visually-rich documents, a lot of research models the visual information extraction task as a computer vision problem, and performs information extraction through semantic segmentation or text box detection. Considering that text information also plays an important role in visual information extraction, the typical framework is to treat document images as a pixel grid and add text features to the visual feature map to obtain a better representation. According to the granularity of textual information, these approaches are developed from character-level to word-level and then to context-level. Chargrid (Katti et al., 2018) uses a convolution-based encoder-decoder network to fuse text information into images by performing one-hot encoding on characters. VisualWordGrid (Kerroumi et al., 2020) implements Wordgrid (Katti et al., 2018) by replacing character-level text information with word-level word2vec features, and fusing visual information to improve the extraction performance. BERTgrid (Denk & Reisswig, 2019) uses BERT to obtain contextual text representation, which further improves the end-to-end accuracy. Based on BERTgrid, ViBERTgrid (Lin et al., 2021) fuses the text features from BERT with the image features from the CNN model, thus obtaining better results.

Since textual information still plays an important role in visually-rich documents, a lot of research work takes information extraction as a special natural language understanding task. Majumder et al.

(2020) generate candidates according to the types of the extracted entities, and has achieved good results in form understanding. TRIE (Zhang et al., 2020) combines text detection and information extraction, allowing two tasks to promote each other to obtain better information extraction results. Wang et al. (2020b) predict the relationship between text fragments through the fusion of three different modalities, and realize the hierarchical extraction for form understanding.

Unstructured visually-rich documents are often composed of multiple adjacent text fragments, so it is also natural to use the Graph Neural Network (GNN) for representation. The text fragments in a document are considered as nodes in the graph, while the relationship between the text fragments can be modeled as edges, so that the entire document can be represented as a graph network. In Section 2.2, we introduced the representative work of GNN for information extraction in visually-rich documents (Liu et al., 2019a). On this basis, there is more research work based on GNN for visual information extraction. Hwang et al. (2020) model the document as a directed graph and extract information from the document through dependency analysis. Riba et al. (2019) use a GNN model to extract tabular information from the invoices. Wei et al. (2020) use Graph Convolutional Networks (GCN) to model the text layout based on the output of the pre-trained models, which improves information extraction. Cheng et al. (2020) achieves better performance in few-shot learning by representing the document as a graph structure and using a graph-based attention mechanism and a CRF model. The PICK (Yu et al., 2021) model introduces a graph that can be learned based on nodes to represent documents, and achieved better performance in receipt understanding.

5.1.3 DOCUMENT IMAGE CLASSIFICATION

Document image classification refers to the task of classifying document images that is essential for business digitalization. RVL-CDIP (Harley et al., 2015) is a representative dataset for this task. The dataset contains 400,000 grayscale images in 16 document image categories. Tabacco-3482 (Kumar et al., 2014) selects a subset of RVL-CDIP for evaluation, which contains 3,482 grayscale document images.

Document image classification is a special subtask of image classification, thus classification models for natural images can also address the problem of document image classification. Afzal et al. (2015) introduce a document image classification method based on CNN for document image classification. To overcome the problem of insufficient samples, they use Alexnet trained with ImageNet as the initialization for model adaptation on document images. Afzal et al. (2017) use GoogLeNet, VGG, ResNet and other successful models from natural images on document images through transfer learning. Through the adjustment of model parameters and data processing, Tensmeyer & Martinez (2017) use the CNN model that can outperform the previous models without transfer learning from natural images. Das et al. (2018) propose a new convolutional network based on different image regions for document image classification. This method classifies different regions of the document separately, and finally merges multiple classifiers of different regions to obtain a significant performance improvement in document image classification. Sarkhel & Nandi (2019) extract features at different levels by introducing a pyramidal multi-scale structure. Dauphinee et al. (2019) obtain the text of the document by performing OCR on the document image, and combine image and text features to further improve the classification performance.

5.1.4 DOCUMENT VISUAL QUESTION ANSWERING

Document Visual Question Answering (VQA) is a high-level understanding task for document images. Specifically, given a document image and a related question, the model needs to give the correct answer to the question based on the given image. A specific example is shown in Figure 5. VQA for documents first appears in the DocVQA dataset (Mathew et al., 2021b), which contains more than 12,000 documents and corresponding 5,000 questions. Later, InfographicVQA (Mathew et al., 2021a) is also proposed, which is a VQA benchmark for infographic images in the documents. As the answers in DocVQA are relatively short and topics are not diverse, some researchers also proposed the VisualMRC (Tanaka et al., 2021) dataset for the document VQA task, which includes long answers with diverse topics.

Different from the traditional VQA task, textual information in document VQA plays a key role in this task, so existing representative methods all take the texts obtained by OCR of document images as the inputs. After the document text is obtained, the VQA task is modeled as different problems

Q: Mention the ZIP code written?
A: 80202

Q: What date is seen on the seal at the top of the letter?
A: 23 sep 1970

Q: Which company address is mentioned on the letter?
A: Great western sugar Co.

(a)

Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?
A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.
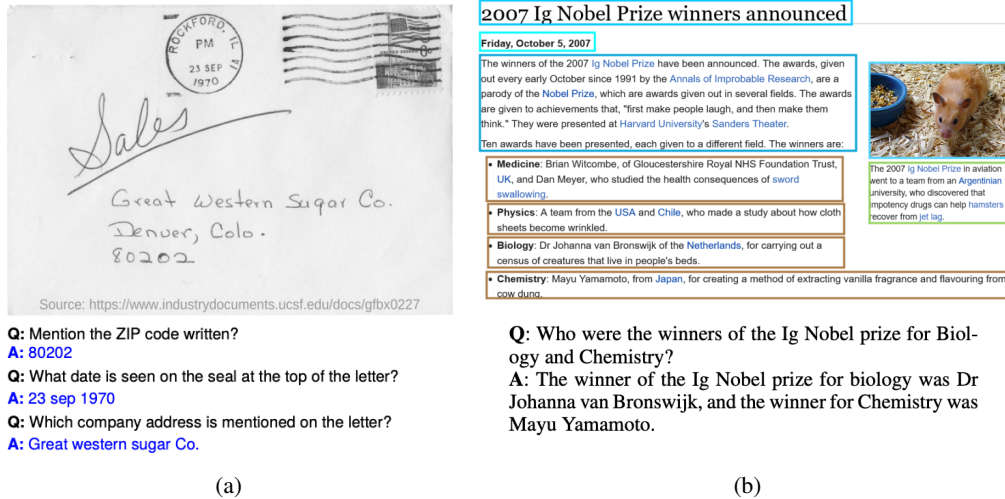
(b)

Figure 5: Examples of Document Visual Question Answering

according to the characteristics of different datasets. For the DocVQA data, most of the answers to questions exist as text fragments in the document text, so mainstream methods have modeled it as the Machine Reading Comprehension (MRC) problem. By providing the model with visual features and document texts, the model extracts text fragments from the given document according to the question as the corresponding answer. For the VisualMRC dataset, the answer to the question usually does not literally appear in the document text fragment and a longer abstract answer is required. Therefore, a feasible method is to use a text generation approach to generate answers to the questions.

5.2 GENERAL-PURPOSE MULTIMODAL PRE-TRAINING

Although the above methods achieve good performance on document understanding tasks, these methods usually have two limitations: 1) The models often rely on limited labeled data, while neglecting a large amount of knowledge in unlabeled data. On one hand, for document understanding tasks such as information extraction, human annotation of data is expensive and time-consuming. On the other hand, due to the extensive use of visually-rich documents in the real world, there are a large number of unlabeled documents, and these large amounts of unlabeled data can be leveraged for self-supervised pre-training. 2) Visually-rich documents not only contain a lot of text information, but also rich layout and visual information. Existing models for specific tasks usually only use pre-trained CV models or NLP models to obtain the knowledge from the corresponding modality due to the limitation of data, and most of the work only uses information from a single modality or simple combination of features rather than the deep fusion. The success of Transformer (Vaswani et al., 2017) in transfer learning proves the importance of deep contextualization for sequence modeling for both NLP and CV problems. Therefore, it is obvious to jointly learn different modalities such as text, layout and visual information in a single framework.

Visually-rich documents mainly involve three modalities: text, layout, and visual information, and these modalities have a natural alignment in visually-rich documents. Therefore, it is vital to model document representations and achieve cross-modal alignment through pre-training. The LayoutLM (Xu et al., 2020) and the subsequent LayoutLMv2 (Xu et al., 2021a) model are proposed as the pioneer work in this research area. In Section 2.3, we introduced LayoutLM, a general pre-trained model for Document AI. Through joint pre-training of text and layout, LayoutLM has achieved significant improvement in a variety of document understanding tasks. On this basis, there is a lot of follow-up research work to improve this framework. LayoutLM does not introduce document visual information in the pre-training process, so the accuracy is not satisfactory on tasks that require strong visual perception such as DocVQA. In response to this problem, LayoutLMv2 (Xu et al., 2021a) integrates visual information into the pre-training process, which greatly improves the visual understanding capability. Specifically, LayoutLMv2 introduces a spatial-aware self-attention

mechanism, and uses visual features as part of the input sequence. For the pre-training objectives, LayoutLMv2 proposes "Text-Image Alignment" and "Text-Image Matching" tasks in addition to Masked Visual-Language Modeling. Through improvements in these two aspects, the model capability to perceive visual information is substantially improved, and it significantly outperforms strong baselines in a variety of downstream Document AI tasks.

Visually-rich documents can be generally divided into two categories. The first one is the fixed-layout documents such as scanned document images and digital-born PDF files, where the layout and style information is pre-rendered and independent of software, hardware, or operating system. This property makes existing layout-based pre-training approaches easily applicable to document understanding tasks. While, the second category is the markup language based documents such as HTML/XML, where the layout and style information needs to be interactively and dynamically rendered for visualization depending on the software, hardware, or operating system. For markup language based documents, the 2D layout information does not exist in an explicit format but usually needs to be dynamically rendered for different devices, e.g. mobile/tablet/desktop, which makes current layout-based pre-trained models difficult to apply. To this end, MarkupLM (Li et al., 2021b) is proposed to jointly pre-train text and markup language in a single framework for markup-based VrDU tasks. Distinct from fixed-layout documents, markup-based documents provide another viewpoint for the document representation learning through markup structures because the 2D position information and document image information cannot be used straightforwardly during the pre-training. Instead, MarkupLM takes advantage of the tree-based markup structures to model the relationship among different units within the document.

Position Information After LayoutLM, much research work has made improvements based on this model framework. One of the main directions is to improve the way of position embeddings. Some work has changed the position encoding represented by embeddings to the sinusoidal functions, such as BROS (Hong et al., 2020) and StructuralLM (Li et al., 2021a). BROS (Hong et al., 2020) uses the sinusoidal function for the absolute position encoding, and at the same time introduced the relative position information of the text through the sinusoidal function in the self-attention mechanism, which improves the model's ability to perceive spatial position. StructuralLM (Li et al., 2021a) shares the same position information in the text block in the absolute position representation, which helps the model understand the text information in the same entity, thereby further improving information extraction.

Visual Information In addition, some research work has made further improvements to optimize and strengthen the vision models. LAMPRET (Wu et al., 2021) provides the model with more visual information to model web documents such as font size, illustrations, etc. which helps to understand rich web data. SelfDoc (Li et al., 2021c) adopts a two-stream structure. For a given visually-rich document image, a pre-trained document entity detection model is first used to identify all semantic units in the document through object detection, then OCR is used to recognize the textual information. For the identified image regions and text sequences, the model uses Sentence-BERT (Reimers & Gurevych, 2019) and Faster-RCNN (Ren et al., 2016) to extract features and encode them as feature vectors. A cross-modal encoder is used for encoding the whole image with a multi-modal representation to serve downstream tasks. DocFormer (Appalaraju et al., 2021) adopts a discrete multi-modal structure and uses position information on each layer to combine text and visual modalities for the self-attention. DocFormer uses ResNet (He et al., 2015) to encode image information to obtain higher resolution image features, and at the same time encodes text information into text embeddings. The position information is added to the image and text information separately and passed to the Transformer layer separately. Under this mechanism, high-resolution image information was obtained while the input sequence was shortened. Meanwhile, different modalities are aligned through position information so that the model could better learn the cross-modal relationship of visually-rich documents.

Pre-training Tasks Moreover, some pre-trained models have designed richer pre-training tasks for different modalities. For example, in addition to the Masked Visual-Language Modeling (MVLM), BROS (Hong et al., 2020) proposes an area-masked language model, which masks all text blocks in a randomly selected area. It can be interpreted as extending the interval mask operation for one-dimensional text in SpanBERT (Joshi et al., 2020) to an interval mask for text blocks in a two-dimensional space. Specifically, the operation consists of the following four steps: (1) randomly

selecting a text block, (2) determining a final area by expanding the area of the text block, (3) determining the text block belonging to the area, (4) masking all the text within block and recover. LAMPRET (Wu et al., 2021) additionally introduces the ordering of web page entities, which allows the model to learn the spatial position by predicting the order of the entity arrangement. At the same time, the model also uses image matching pre-training by removing images in the webpage and matching through retrieval. This also improves the model's ability to understand the semantics of multimodal information. The "Cell Position Classification" task proposed by StructuralLM (Li et al., 2021a) models the relative spatial position of the text block in the document. Given a set of scanned files, this task aims to predict the location of text blocks in the file. First, a visually-rich document is divided into N regions of the same size. Then, the model calculates the area to which the text block belongs through the two-dimensional position of the center of the text block. SelfDoc (Li et al., 2021c) and DocFormer (Appalaraju et al., 2021) also introduce new pre-training tasks along with the improvements of image inputs. SelfDoc masks and predicts the image features to better learn the visual information. DocFormer introduces a decoder to reconstruct image information. In this case, the task is similar to the image reconstruction of an autoencoder, but it contains multimodal features such as texts and positions. With the help of joint image and text pre-training, image reconstruction requires the deep fusion of texts and images, which strengthens the interaction between different modalities.

Initialization   Regarding model initialization, some approaches used the existing powerful pre-trained language models to further improve their performance, while also expanding the capabilities of the pre-trained models. For example, LAMBERT (Garncarek et al., 2020) achieves better performance by using RoBERTa (Liu et al., 2019b) as the pre-training initialization. In addition to language understanding, some models focus on extending the language generation capabilities of the models. A common practice is to use the encoder-decoder models for initialization. TILT (Powalski et al., 2021) introduces the layout encoding layer into the pre-trained T5 (Raffel et al., 2020) model and combined document data for pre-training, so that the model can handle the generation tasks in Document AI. LayoutT5 and LayoutBART (Tanaka et al., 2021) introduce text position encoding on top of T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) models in the fine-tuning stage for document VQA to help the model understand questions and generate answers better.

Multilingual   Although these models have been successfully applied on English documents, document understanding tasks are also important for non-English speaking worlds. LayoutXLM (Xu et al., 2021b) is the first research work that carries out multilingual pre-training on visually-rich documents for other languages. Based on the model structure of LayoutLMv2, LayoutXLM expands the language support of LayoutLM by using 53 languages for pre-training. Compared with the cross-language models for plain text, LayoutXLM has obvious advantages in the language expansion capability for visually-rich documents, which proves that cross-lingual pre-training not only works on pure NLP tasks, but also effective for cross-lingual Document AI tasks.

6   CONCLUSION AND FUTURE WORK

Automated information processing is the foundation and prerequisite for digital transformation. Nowadays, there are increasingly higher requirements for processing power, processing speed, and processing accuracy. Taking the business field as an example, electronic business documents cover a large amount of complicated information such as purchase receipts, industry reports, business emails, sales contracts, employment agreements, commercial invoices, and personal resumes. The Robotic Process Automation (RPA) industry has been created in this background, using AI technology to help a large number of people free from complicated electronic document processing tasks, meanwhile improving productivity through a series of supporting automation tools. One of the key cores of RPA is the Document AI technique. In the past 30 years, document analysis has mainly gone through three stages, from the early-stage heuristic rules, to the statistical machine learning, and recently deep learning methods, which greatly advances analysis performance and accuracy. At the same time, we have also observed that the large-scale self-supervised general document-level pre-trained model represented by LayoutLM has also received more and more attention and usage, and has gradually become the basic unit for building more complex algorithms. There are also quite a few follow-up research work that emerged recently, which accelerates development of the Document AI.

For future research, in addition to the multi-page/cross-page problems, uneven quality of training data, weak multi-task relevance, and few-shot and zero-shot learning, we also need to pay special attention to the relationship between OCR with Document AI tasks, since the input of Document AI applications usually comes from automatic OCR models. The accuracy of text recognition often has a great impact on downstream tasks. In addition, how to combine Document AI technology with existing human knowledge especially manual document processing skills, is an interesting research topic worth exploring in the future.

REFERENCES

Abdelrahman Abdallah, Alexander Berendeyev, Islam Nuradin, and Daniyar Nurseitov. Tncr: Table net detection and classification dataset. *arXiv preprint arXiv:2106.15322*, 2021.

Muhammad Zeshan Afzal, Samuele Capobianco, Muhammad Imran Malik, Simone Marinai, Thomas M Breuel, Andreas Dengel, and Marcus Liwicki. Deepdocclassifier: Document classification with deep convolutional neural network. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pp. 1111–1115. IEEE, 2015.

Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pp. 883–888. IEEE, 2017.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*, 2021.

Micheal Baechler and Rolf Ingold. Multi resolution layout analysis of medieval manuscripts using dynamic mlp. In *2011 International Conference on Document Analysis and Recognition*, pp. 1185–1189. IEEE, 2011.

Henry S Baird, Susan E Jones, and Steven J Fortune. Image segmentation by shape-directed covers. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pp. 820–825. IEEE, 1990.

Anukriti Bansal, Gaurav Harit, and Sumantra Dutta Roy. Table extraction from document images using fixed point model. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, pp. 1–8, 2014.

Itay Bar-Yosef, Nate Hagbi, Klara Kedem, and Itshak Dinstein. Line segmentation for degraded handwritten historical documents. In *2009 10th International Conference on Document Analysis and Recognition*, pp. 1161–1165. IEEE, 2009.

Philippine Barlas, Sébastien Adam, Clément Chatelain, and Thierry Paquet. A typed and handwritten text block segmentation system for heterogeneous and complex documents. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pp. 46–50. IEEE, 2014.

Galal M Binmakhashen and Sabri A Mahmoud. Document layout analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36, 2019.

Dan S Bloomberg. Multiresolution morphological approach to document image analysis. In *Proc. of the international conference on document analysis and recognition, Saint-Malo, France*, 1991.

Syed Saqib Bukhari, Faisal Shafait, and Thomas M Breuel. Script-independent handwritten textlines segmentation using active contours. In *2009 10th International Conference on Document Analysis and Recognition*, pp. 446–450. IEEE, 2009.

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*, pp. 183–190, 2010.

Syed Saqib Bukhari, Thomas M Breuel, Abedelkadir Asi, and Jihad El-Sana. Layout analysis for arabic historical document images using machine learning. In *2012 International Conference on Frontiers in Handwriting Recognition*, pp. 639–644. IEEE, 2012.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

Jin Chen and Daniel Lopresti. Table detection in noisy off-line handwritten documents. In *2011 International Conference on Document Analysis and Recognition*, pp. 399–403. IEEE, 2011.

Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pp. 965–970. IEEE, 2017.

Lu Chen, Xingyu Chen, Zihan Zhao, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension, 2021.

Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. One-shot text field labeling using attention and belief propagation for structure information extraction. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 340–348, 2020.

Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan K Parui. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3180–3185. IEEE, 2018.

Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. Modular multimodal architecture for document classification. *arXiv preprint arXiv:1912.04376*, 2019.

Timo I Denk and Christian Reisswig. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*, 2019.

Harsh Desai, Pratik Kayal, and Mayank Singh. Tablex: A benchmark dataset for structure and content information extraction from scientific tables, 2021.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

Haoyu Dong, Shijie Liu, Shi Han, Zhouyu Fu, and Dongmei Zhang. Tablesense: Spreadsheet table detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 69–76, 2019.

Ana Costa e Silva. Learning rich hidden markov models in document analysis: Table location. In *2009 10th International Conference on Document Analysis and Recognition*, pp. 843–847. IEEE, 2009.

Floriana Esposito, Donato Malerba, Giovanni Semeraro, Enrico Annese, and Giovanna Scafuro. An experimental page layout recognition system for office document automatic classification: an integrated approach for inductive generalization. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pp. 557–562. IEEE, 1990.

Floriana Esposito, Stefano Ferilli, Teresa MA Basile, and Nicola Di Mauro. Machine learning for digital document processing: From layout analysis to metadata extraction. In *Machine learning in document analysis and recognition*, pp. 105–138. Springer, 2008.

Jing Fang, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pp. 445–449. IEEE, 2012.

James L Fisher, Stuart C Hinds, and Donald P D'Amato. A rule-based system for document image segmentation. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pp. 567–572. IEEE, 1990.

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)*, pp. 1510–1515, 2019. doi: 10.1109/ICDAR.2019.00243.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. Lambert: Layout-aware (language) modeling for information extraction. *arXiv preprint arXiv:2002.08087*, 2020.

Max C. Göbel, Tamir Hassan, Ermelinda Oro, and G. Orsi. Icdar 2013 table competition. *2013 12th International Conference on Document Analysis and Recognition*, pp. 1449–1453, 2013.

Tobias Grüning, Gundram Leifert, Tobias Strauß, Johannes Michael, and Roger Labahn. A two-stage method for text line detection in historical documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(3):285–302, 2019.

Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*, 2019.

He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Eaten: Entity-aware attention for single shot visual text extraction, 2019.

Qiang Hao, Rui Cai, Yanwei Pang, and Lei Zhang. From one tree to a forest: A unified solution for structured web data extraction. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pp. 775–784, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307574. doi: 10.1145/2009916.2010020. URL https://doi.org/10.1145/2009916.2010020.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*, 2020.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model for understanding texts in document. 2020.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sep 2019. doi: 10.1109/icdar.2019.00244. URL http://dx.doi.org/10.1109/ICDAR.2019.00244.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. Spatial dependency parsing for semi-structured document information extraction. *arXiv preprint arXiv:2005.00642*, 2020.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Thotreingam Kasar, Philippine Barlas, Sebastien Adam, Clément Chatelain, and Thierry Paquet. Learning to detect tables in scanned document images using line information. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 1185–1189. IEEE, 2013.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4459–4469, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1476. URL `https://www.aclweb.org/anthology/D18-1476`.

Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. Visualwordgrid: Information extraction from scanned documents using a multimodal approach. *arXiv preprint arXiv:2010.02358*, 2020.

Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382, 1998.

J. Kumar, Peng Ye, and D. Doermann. Structural similarity for document image classification and retrieval. *Pattern Recognit. Lett.*, 43:119–126, 2014.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL `https://aclanthology.org/2020.acl-main.703`.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*, 2021a.

Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. Markuplm: Pre-training of text and markup language for visually-rich document understanding, 2021b.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1918–1925, Marseille, France, May 2020a. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.236`.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 949–960, Barcelona, Spain (Online), December 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.82. URL `https://aclanthology.org/2020.coling-main.82`.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5652–5660, 2021c.

Quannan Li, Jingdong Wang, David Wipf, and Zhuowen Tu. Fixed-point model for structured labeling. In *International conference on machine learning*, pp. 214–221. PMLR, 2013.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents. *arXiv preprint arXiv:2105.11672*, 2021.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, pp. 21–37, 2016. ISSN 1611-3349. doi: 10.1007/978-3-319-46448-0_2. URL `http://dx.doi.org/10.1007/978-3-319-46448-0_2`.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pp. 32–39, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-2005. URL `https://aclanthology.org/N19-2005`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019b.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pp. 6495–6504, 2020.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021a.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021b.

Ajoy Mondal, Peter Lipps, and CV Jawahar. Iiit-ar-13k: a new dataset for graphical object detection in documents. In *International Workshop on Document Analysis Systems*, pp. 216–230. Springer, 2020.

George Nagy and Sharad C Seth. Hierarchical representation of optically scanned documents. 1984.

Lawrence O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 15(11):1162–1173, 1993.

Masayuki Okamoto and Makoto Takahashi. A hybrid page segmentation method. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pp. 743–746. IEEE, 1993.

Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 7–12. IEEE, 2018.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. 2019.

David Pinto, Andrew McCallum, Xing Wei, and W Bruce Croft. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 235–242, 2003.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. Going full-tilt boogie on document understanding with text-image-layout transformer. *arXiv preprint arXiv:2102.09550*, 2021.

Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 572–573, 2020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Sheikh Faisal Rashid, Abdullah Akmal, Muhammad Adnan, Ali Adnan Aslam, and Andreas Dengel. Table recognition in heterogeneous documents using machine learning. In *2017 14th IAPR International conference on document analysis and recognition (ICDAR)*, volume 1, pp. 777–782. IEEE, 2017.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL `https://aclanthology.org/D19-1410`.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 122–127. IEEE, 2019.

Takashi Saitoh, Michiyoshi Tachikawa, and Toshifumi Yamaai. Document image segmentation and text area ordering. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pp. 323–329. IEEE, 1993.

Ritesh Sarkhel and Arnab Nandi. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *28th International Joint Conference on Artificial Intelligence (IJCAI), 2019*, 2019.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pp. 1162–1167, 2017. doi: 10.1109/ICDAR.2017.192.

Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pp. 113–120, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587738. doi: 10.1145/1815330.1815345. URL `https://doi.org/10.1145/1815330.1815345`.

Zhixin Shi and Venu Govindaraju. Line separation for complex document images using fuzzy run-length. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pp. 306–312. IEEE, 2004.

Shoaib Ahmed Siddiqui, Muhammad Imran Malik, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. Decnt: Deep deformable cnn for table detection. *IEEE Access*, 6:74151–74161, 2018.

Raymond W Smith. Hybrid page layout analysis via tab-stop detection. In *2009 10th International Conference on Document Analysis and Recognition*, pp. 241–245. IEEE, 2009.

Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards a universal dataset and metrics for training and evaluating table extraction models, 2021.

Carlos Soto and Shinjae Yoo. Visual detection with context for document layout analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3462–3468, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1348. URL `https://www.aclweb.org/anthology/D19-1348`.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts, 2021.

Jonathan Stray and Stacey Svetlichnaya. Project deepform: Extract information from documents, 2020. URL `https://wandb.ai/deepform/political-ad-extraction/benchmark`.

Don Sylwester and Sharad Seth. A trainable, single-pass algorithm for column segmentation. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pp. 615–618. IEEE, 1995.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. *arXiv preprint arXiv:2101.11272*, 2021.

Chris Tensmeyer and Tony Martinez. Analysis of convolutional neural networks for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pp. 388–393. IEEE, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Matheus Palhares Viana and Dário Augusto Borges Oliveira. Fast cnn-based document layout analysis. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1173–1180, 2017.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. Towards robust visual information extraction in real world: New dataset and novel solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2738–2745, 2021a.

Yalin Wang, Robert Haralick, and Ihsin T Phillips. Improvement of zone content classification by using background analysis. In *Fourth IAPR International Workshop on Document Analysis Systems.(DAS2000)*. Citeseer, 2000.

Yalin Wang, Ihsin T Phillips, and Robert M Haralick. Table detection via probability optimization. In *International Workshop on Document Analysis Systems*, pp. 272–282. Springer, 2002.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. Structure-aware pre-training for table understanding with tree-based transformers. *arXiv preprint arXiv:2010.12537*, 2020a.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. *arXiv preprint arXiv:2010.11685*, 2020b.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. Layoutreader: Pre-training of text and layout for reading order detection, 2021b.

Yalin Wangt, Ihsin T Phillipst, and Robert Haralick. Automatic table ground truth generation and a background-analysis-based table structure extraction method. In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pp. 528–532. IEEE, 2001.

Hao Wei, Micheal Baechler, Fouad Slimane, and Rolf Ingold. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 1220–1224. IEEE, 2013.

Mengxi Wei, Yifan He, and Qiong Zhang. Robust layout-aware ie for visually rich documents with pre-trained language models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2367–2376, 2020.

Christoph Wick and Frank Puppe. Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 287–292. IEEE, 2018.

Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl. Document analysis system. *IBM journal of research and development*, 26(6):647–656, 1982.

Chung-Chih Wu, Chien-Hsing Chou, and Fu Chang. A machine-learning approach for analyzing document layout structures with two reading orders. *Pattern recognition*, 41(10):3200–3213, 2008.

Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky. Lampret: Layout-aware multimodal pretraining for document understanding. *arXiv preprint arXiv:2104.08405*, 2021.

Yi Xiao and Hong Yan. Text region extraction in a document image based on the delaunay tessellation. *Pattern Recognition*, 36(3):799–809, 2003.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2579–2591, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.201. URL `https://aclanthology.org/2021.acl-long.201`.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403172. URL `https://doi.org/10.1145/3394486.3403172`.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding, 2021b.

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural network, 2017a.

Xiaowei Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4342–4351, 2017b.

Antonio Jimeno Yepes, Xu Zhong, and Douglas Burdick. Icdar 2021 competition on scientific literature parsing, 2021.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. Pick: Processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4363–4370. IEEE, 2021.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. Trie: End-to-end text reading and information extraction for document understanding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1413–1422, 2020.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. *arXiv preprint arXiv:1911.10683*, 2019a.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1015–1022. IEEE, Sep. 2019b. doi: 10.1109/ICDAR.2019.00166.