End-to-End Document Layout Analysis using Deep Learning for the DocBank Dataset

Gajendra Shravan Mali (1037955)

Research Proposal for

Master of Science in Machine Learning & Artificial Intelligence

Liverpool John Moores University & upGrad

July 2023

Abstract

Document layout analysis (DLA) is the task of extracting structural information from documents, such as the layout of text, tables, and images. In recent years, deep learning has emerged as a powerful tool for DLA, and several large-scale datasets have been created to facilitate the development of deep learning models for this task. This paper presents an end-to-end deep learning approach to DLA that combines the YOLOv8 object detection model with the PaddleOCR optical character recognition (OCR) framework. YOLOv8 is a state-of-the-art object detection model that can be used to quickly and accurately detect text in documents. PaddleOCR is an OCR framework that can be used to recognize text in images. The proposed approach first uses YOLOv8 to detect text in a document. The detected text is then passed to PaddleOCR, which is used to recognize the text and extract the document layout information. This approach leverages the strengths of both YOLOv8 and PaddleOCR to achieve accurate and comprehensive document layout analysis.

Table of contents

1. Introduction

1.1 Background information on document layout analysis

Document layout analysis refers to the process of extracting structural information from documents in order to understand their organization and arrangement. It involves segmenting a document into its constituent parts, such as paragraphs, headers, footers, captions, and images, and determining their spatial relationships. The layout of a document provides important cues for understanding its content and context. By analyzing the document layout, it becomes possible to discern the logical structure, hierarchy, and relationships between different elements within the document. This information is crucial for various document processing tasks, such as information extraction, document summarization, content-based retrieval, and document understanding.

Traditional approaches to document layout analysis often relied on rule-based methods and handcrafted features. These methods utilized heuristics and predefined rules to detect and classify document elements based on properties like font size, indentation, and whitespace. While these approaches worked reasonably well for structured documents with consistent layouts, they struggled with documents that exhibited complex layouts, variations in formatting, and heterogeneous content. With the advent of deep learning techniques, there has been a paradigm shift in document layout analysis. "Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)", have shown remarkable performance in various computer vision and natural language processing tasks. These models excel at learning complex patterns and representations directly from raw data, making them well-suited for the analysis of document layouts.

Deep learning approaches for document layout analysis typically involve training models on large datasets of annotated documents. The models learn to recognize and classify different document elements based on the visual features present in the document images. This data-driven approach allows for more flexibility and adaptability to diverse document layouts and content. The availability of benchmark datasets, such as the DocBank dataset, has further fueled advancements in deep learning-based document layout analysis. The DocBank dataset comprises a large collection of documents from various domains, along with ground truth annotations for different layout elements. This dataset enables researchers to develop and evaluate algorithms for end-to-end document layout analysis.

In this paper, we propose an end-to-end document layout analysis framework using deep learning techniques on the DocBank dataset. Our goal is to develop a robust and accurate system that can automatically analyze the layout of documents and extract valuable structural information. By leveraging the power of deep learning models, we aim to overcome the limitations of traditional methods and achieve state-of-the-art performance in document layout analysis. The remainder of this paper is organized as follows: Section II provides a comprehensive literature review on document layout analysis, including traditional approaches and recent advancements in deep learning methods. Section III describes the data preprocessing steps, including an in-depth discussion of the DocBank dataset. Section IV presents our proposed methodology for end-to-end document layout analysis. Section V discusses the experimental setup, evaluation metrics, and comparison with baseline models.

1.2 Importance of document layout analysis in various applications

Document layout analysis plays a crucial role in a wide range of applications that involve the processing and understanding of documents. The analysis of document layout provides valuable insights into the organization and structure of textual and visual elements within a document. Here are some key applications where document layout analysis is of utmost importance:

- **Optical Character Recognition (OCR):** Document layout analysis is a fundamental step in OCR systems. By accurately identifying and segmenting text regions, OCR algorithms can extract and recognize the text content of documents. The layout information helps OCR systems differentiate between text and non-text regions, handle variations in font sizes and styles, and improve the accuracy of text recognition.
- **Information Extraction**: Extracting relevant information from documents is a common task in various domains, such as finance, legal, healthcare, and research. Document layout analysis assists in locating and isolating specific sections or fields of interest, such as names, dates, addresses, or tables, which can then be extracted for further processing or analysis.
- **Document Understanding:** Document layout analysis plays a vital role in understanding the semantic structure and hierarchy of a document. By identifying and categorizing different layout elements, such as headings, subheadings, paragraphs, bullet points, and captions, systems can gain insights into the document's organization and meaning. This understanding enables higher-level tasks like document summarization, document classification, and topic modeling.
- **Content-Based Document Retrieval:** Document layout analysis facilitates efficient retrieval of relevant documents based on their visual appearance. By indexing and searching documents based on their layout features, users can retrieve documents with similar structures or visual arrangements. This is particularly useful in scenarios where the visual layout is a key criterion for document retrieval, such as in architectural plans, magazine layouts, or web page analysis.
- **Document Visualization and Accessibility**: Analyzing document layout is essential for creating visually appealing and accessible document representations. By understanding the layout structure, systems can generate alternative document

representations, such as reflowable formats, audio descriptions, or braille versions, to cater to individuals with visual impairments or different reading preferences.

- **Document Forensics:** Document layout analysis plays a role in forensic analysis, where the authenticity, integrity, or origin of a document needs to be verified. By analyzing the layout properties, such as watermark patterns, document templates, or inconsistencies in page alignment, experts can identify potential fraud or tampering.

The importance of document layout analysis lies in its ability to uncover the structural and semantic information embedded within documents. By accurately analyzing the layout, systems can better interpret the content, extract relevant information, facilitate efficient retrieval, and enable a range of downstream applications. With the advancements in deep learning techniques, document layout analysis has witnessed significant progress, leading to improved performance, and expanding the possibilities for document processing and understanding.

1.3 Introduction to the DocBank dataset

The DocBank dataset is a comprehensive benchmark dataset specifically designed for document layout analysis tasks. It serves as a valuable resource for researchers and practitioners in the field to develop and evaluate algorithms for end-to-end document layout analysis using deep learning techniques. The DocBank dataset encompasses a diverse collection of documents from various domains, including scientific papers, legal documents, news articles, and reports. It captures the wide range of layouts and design choices encountered in real-world documents. The dataset is carefully annotated with ground truth labels for different layout elements, providing the necessary training and evaluation data for document layout analysis tasks.

The annotations in the DocBank dataset cover a broad set of layout elements, such as titles, headings, paragraphs, captions, footnotes, figures, tables, and equations. Each layout element is associated with its spatial coordinates or bounding box, indicating its exact position within the document. Additionally, the dataset also includes hierarchical structural information, such as section headings and subsections, capturing the document's organization. The DocBank dataset is designed to support end-to-end document layout analysis tasks, where the goal is to recognize and classify different layout elements within a document. This includes tasks like text detection, text recognition, semantic segmentation, and document structure analysis. The dataset allows researchers to train and evaluate deep learning models that can understand the complex spatial relationships between different layout elements in a document.

The availability of the DocBank dataset addresses a significant gap in the research community by providing a standardized and comprehensive benchmark for document layout analysis. It enables fair comparisons between different algorithms and promotes the development of more accurate and robust systems for document layout analysis. The DocBank dataset also encourages advancements in the field by stimulating research in areas such as multi-domain document analysis, handling complex layouts, dealing with low-resource scenarios, and exploring the integration of contextual information for improved layout understanding.

In this paper, we leverage the DocBank dataset to develop and evaluate our end-to-end document layout analysis approach using deep learning techniques. By utilizing this dataset, we aim to demonstrate the effectiveness and applicability of our proposed methodology in real-world document analysis scenarios. The following sections of this paper will detail our approach, experimental setup, results, and discussions, highlighting the insights gained from utilizing the DocBank dataset and the implications for the field of document layout analysis.

1.4 Statement of the problem and research objectives

The problem addressed in this research is the accurate and robust analysis of document layout using deep learning techniques. Document layout analysis is a complex task due to the variability in document designs, content types, and formatting choices. Traditional rule-based methods often struggle to handle such diversity, leading to suboptimal performance. Therefore, there is a need for advanced techniques that can automatically and effectively analyze the layout of documents, enabling accurate interpretation and extraction of relevant information.

1.5 Research objectives

The primary objective of this research is to develop an end-to-end document layout analysis system using deep learning methods, specifically tailored for the DocBank dataset. The system aims to accurately identify and classify different layout elements within documents, such as titles, headings, paragraphs, figures, and tables. Additionally, it aims to capture the hierarchical structure and spatial relationships between these elements, providing a comprehensive understanding of the document layout.

The specific research objectives can be summarized as follows:

- **Preprocessing and Data Preparation:** Conduct preprocessing steps on the DocBank dataset, including data cleaning, normalization, and partitioning into training, validation, and test sets. Prepare the data for subsequent deep learning model training and evaluation.
- **Deep Learning Model Development**: Design and implement a deep learning architecture suitable for end-to-end document layout analysis. This model should effectively learn the visual features and spatial relationships of layout elements from the document images. It should handle the multi-class classification problem of

identifying various layout elements and also address the bounding box regression task for accurate localization.

- **Training and Optimization:** Train the deep learning model using the prepared dataset. Optimize the model's hyperparameters, such as learning rate, batch size, and network architecture, to achieve the best performance on the task of document layout analysis. Employ appropriate loss functions and evaluation metrics to guide the training process.
- **Performance Evaluation**: Evaluate the developed model's performance on the DocBank dataset. Assess its accuracy, precision, recall, and other relevant metrics for layout element classification and bounding box regression. Compare the performance of the proposed approach with existing baseline models or methods to showcase its effectiveness.
- **Analysis and Discussion**: Analyze the results obtained from the experiments and provide a detailed discussion of the strengths, weaknesses, and limitations of the proposed approach. Investigate the impact of different hyperparameters, dataset characteristics, and document types on the model's performance. Discuss potential areas for improvement and future research directions in document layout analysis.

By addressing these research objectives, we aim to contribute to the advancement of document layout analysis using deep learning techniques, specifically focusing on the challenges and opportunities presented by the DocBank dataset. The developed system can serve as a foundation for more sophisticated document processing applications, enhancing the accuracy and efficiency of information extraction, document understanding, and content-based retrieval tasks.

1.6 Overview of the proposed approach and its significance

Our proposed approach for end-to-end document layout analysis utilizes deep learning techniques to accurately analyze the layout of documents in the DocBank dataset. The approach consists of several key steps, including data preprocessing, deep learning model development, training and optimization, and performance evaluation. By leveraging the power of deep learning models, our approach aims to achieve accurate classification of layout elements and precise localization through bounding box regression.

The significance of the proposed approach:

- **Improved Accuracy:** Deep learning models have demonstrated remarkable capabilities in capturing complex patterns and representations from raw data. By utilizing these models for document layout analysis, we expect to achieve higher accuracy in identifying and classifying layout elements within documents. This accuracy translates into more reliable and precise extraction of information from documents, benefiting downstream applications.
- **Robustness to Document Variability:** The DocBank dataset encompasses a wide variety of document layouts, representing diverse domains and content types. Our

proposed approach aims to be robust to this variability, enabling effective analysis and interpretation of documents with different structures and formatting choices. This robustness ensures the generalizability and applicability of the approach to real-world document analysis scenarios.

- **End-to-End Analysis:** Our approach focuses on end-to-end document layout analysis, where the system learns to recognize and classify layout elements directly from raw document images. This holistic approach eliminates the need for manual feature engineering or rule-based methods, allowing the system to learn and adapt to the complexities and variations present in document layouts. The end-to-end nature of the approach simplifies the analysis pipeline and enhances efficiency.
- **Utilization of the DocBank Dataset**: The utilization of the DocBank dataset is a significant strength of our approach. The dataset provides a rich and diverse collection of annotated documents, enabling comprehensive training, validation, and evaluation of the deep learning model. By leveraging this dataset, we ensure the relevance and reliability of our results, while also contributing to the advancement of document layout analysis research.
- **Potential for Real-World Applications**: Accurate document layout analysis has broad implications for various real-world applications. By developing an effective and robust system, we enable more accurate information extraction, improved document understanding, and enhanced content-based retrieval. This can benefit domains such as academia, law, journalism, and administration, where efficient processing and understanding of documents are critical.

The proposed approach, with its focus on accuracy, robustness, end-to-end analysis, and utilization of the DocBank dataset, holds great significance in advancing the field of document layout analysis. By demonstrating the effectiveness and applicability of our approach, we contribute to the development of more sophisticated document processing systems and pave the way for further research and innovation in this domain.

2. Literature Review

2.1 Overview of existing document layout analysis techniques

Document layout analysis is a crucial task in document processing, enabling efficient information extraction and understanding. Several techniques have been proposed to tackle the challenges of document layout analysis, utilizing deep learning and other approaches. In this section, we provide an overview of some notable techniques in the field.

- **"DocBank: A Benchmark Dataset for Document Layout Analysis" by Feng et al. (2020)** *(Li et al., n.d.)*

This paper introduces the DocBank dataset, which serves as a benchmark dataset for document layout analysis. It emphasizes the importance of the dataset in advancing research in the field and provides insights into the dataset creation process. The availability of this dataset enables researchers to develop and evaluate novel techniques for end-to-end document layout analysis.

- **"LayoutLMv2: Multi-modal Pre-training for Visually rich Document Understanding" by Yang Xu et al. (2022)** *(Xu et al., 2020a; b; [2012.14740] LayoutLMv2: Multi-modal Pre-training for Visually Rich Document Understanding, 2023)*

This paper presents LayoutLMv2, a multi-modal pre-training approach for visually rich document understanding. The model combines textual and visual information to capture both textual semantics and visual layout structures. The experimental results demonstrate the effectiveness of jointly considering text and layout in processing visually complex documents, highlighting the potential of multi-modal approaches in document layout analysis.

- **"DiT: Self-supervised Pre-training for Document Image Transformer" by Junlong Li et al. (2022)** *(Li et al., 2022; DiT: Self-supervised Pre-training for Document Image Transformer | Proceedings of the 30th ACM International Conference on Multimedia, 2023)*

This paper introduces DiT, a self-supervised pre-training method for document image understanding. By leveraging unlabeled document images, DiT learns to capture hierarchical structure and semantic information. The proposed method demonstrates the potential of self-supervised learning in improving document understanding tasks, including layout analysis.

- **"Document Layout Analysis via Dynamic Residual Feature Fusion" by Ziling Hu et al. (2021)** *(Wu et al., 2021a; b; [2104.02874] Document Layout Analysis via Dynamic Residual Feature Fusion, 2023; Document Layout Analysis via Dynamic Residual Feature Fusion | IEEE Conference Publication | IEEE Xplore, 2023)*

This paper proposes a method for document layout analysis that incorporates dynamic residual feature fusion. By combining multi-scale features and dynamically fusing them using residual connections, the approach captures both local and global contextual information in document layouts. The experimental results demonstrate improved performance on various layout analysis tasks, showcasing the effectiveness of the proposed method.

- **"BINYAS: A Complex Document Layout Analysis System" by Showmik Bhowmik et al. (2021)** *(Bhowmik et al., 2021; BINYAS: a complex document layout analysis system | SpringerLink, 2023)*

This paper presents BINYAS, a comprehensive system designed to analyze complex document layouts. BINYAS integrates layout understanding, component extraction, and semantic interpretation techniques to accurately analyze and extract information from documents. The evaluation results and application scenarios highlight the effectiveness and practicality of BINYAS in various document-centric tasks.

These techniques represent different approaches to document layout analysis, including benchmark dataset creation, multi-modal pre-training, self-supervised learning, feature fusion, and comprehensive analysis systems. They contribute to the advancement of the field by addressing the challenges of layout understanding and enabling more accurate and efficient processing of documents. By building upon these techniques, our proposed approach aims to further improve end-to-end document layout analysis using deep learning methods, specifically tailored for the DocBank dataset.

2.2 Deep learning methods in document layout analysis

Deep learning methods have shown significant promise in various fields, including document layout analysis. In this section, we explore some notable deep-learning methods that have been employed for document layout analysis tasks.

- **Convolutional Neural Networks (CNNs):**

CNNs have been widely used in document layout analysis for tasks such as text detection, element classification, and semantic segmentation. These networks are adept at capturing local features and patterns within document images. They typically employ convolutional layers to extract hierarchical representations, followed by pooling layers for down sampling and fully connected layers for classification. CNN-based approaches have achieved impressive results in accurately identifying layout elements in documents.

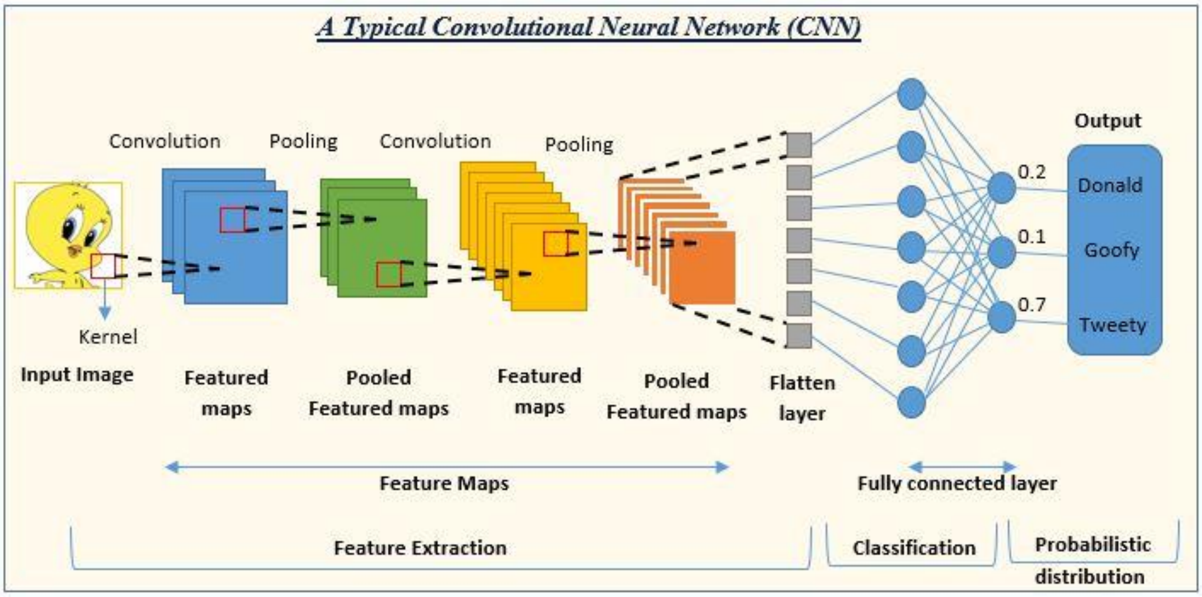


Figure 2.2.1 Typical Convolution Neural Network (CNN) (Convolutional Neural Network: An Overview, 2023)

- **Recurrent Neural Networks (RNNs):**

RNNs, particularly Long Short-Term Memory (LSTM) networks, have been employed to capture sequential dependencies in document layout analysis. They excel at modeling temporal and contextual information, which is crucial for understanding the structural relationships between different layout elements. RNNs have been utilized for tasks such as sequence labeling, structural analysis, and document parsing. They have shown effectiveness in capturing the sequential nature of document layouts.

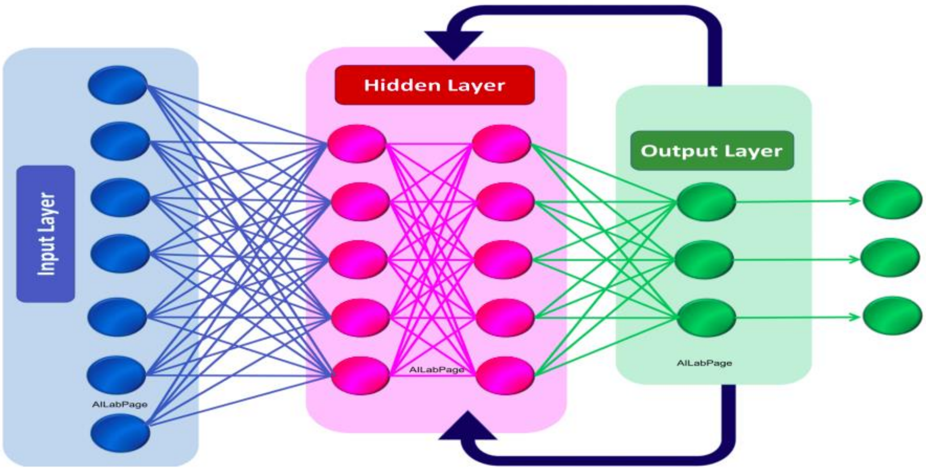


Figure 2.2.2 Recurrent Neural Network (Recurrent Neural Network. So why do we need a recurrent neural… | by Devanshi | DataDrivenInvestor, 2023)

- **Transformer Networks:**

Transformer networks have gained significant attention in various natural language processing tasks, including document layout analysis. These networks utilize self-attention mechanisms to capture global dependencies and relationships within the document. Transformer-based models, such as LayoutLMv2, have been proposed to jointly model text and layout structures, achieving improved performance in visually rich document understanding. The ability of transformer networks to capture long-range dependencies makes them well-suited for analyzing complex document layouts.

- **Encoder-Decoder Architectures:**

Encoder-decoder architectures, such as the U-Net architecture, have been employed for tasks like semantic segmentation and document structure extraction. These architectures consist of an encoder network that encodes the input document image into high-level feature representations and a decoder network that generates predictions or segmentations based on the encoded features. Encoder-decoder architectures have shown effectiveness in capturing both local and global contextual information, enabling accurate layout analysis.

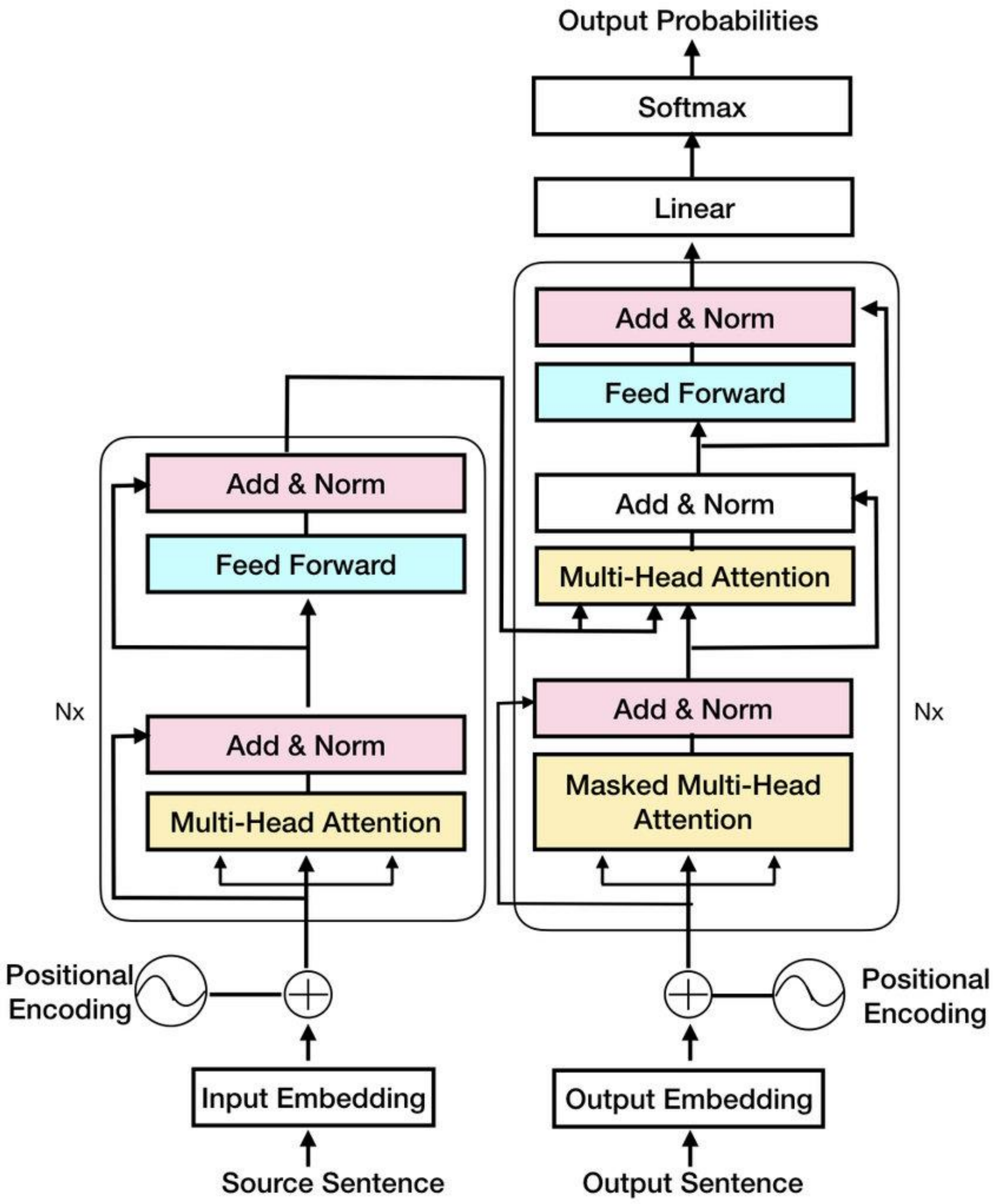
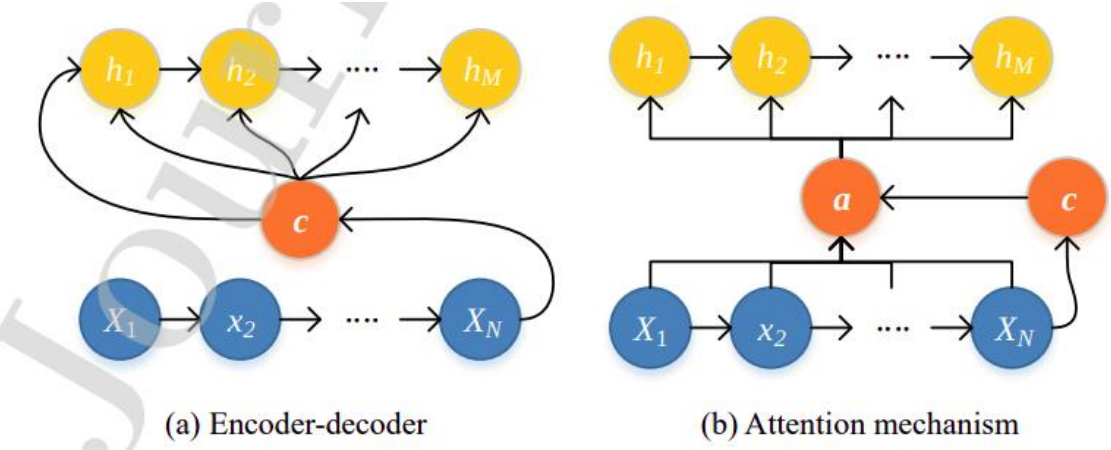


Figure 2.2.3 Encoder Decoder Architecture(Transformer Encoder-Decoder architecture, taken from Vaswani et al. [9]... | Download Scientific Diagram, 2023)

- **Attention Mechanisms:**

At (Chen et al., 2020) attention mechanisms have been incorporated into deep learning models for document layout analysis to selectively focus on relevant regions or elements within the document. Self-attention mechanisms allow the model to attend to different parts of the document with varying importance, improving the accuracy of element classification and localization. Attention-based models have demonstrated their ability to effectively capture the salient features of layout elements.

These deep learning methods have revolutionized document layout analysis by enabling accurate and automated analysis of document structures. By leveraging their ability to capture complex patterns, local and global dependencies, and sequential information, these methods have advanced the field and paved the way for more sophisticated and robust document layout analysis systems. In our proposed approach, we will explore and adapt appropriate deep learning methods to tackle the challenges of end-to-end document layout analysis using the DocBank dataset.



(a) Encoder-decoder

(b) Attention mechanism

Figure 2.2.4 Encoder-decoder & Attention mechanism (Chen et al., 2020; A novel deep learning method based on attention mechanism for bearing remaining useful life prediction - ScienceDirect, 2023)

2.3 Review of relevant studies using deep learning for similar tasks

In recent years, deep learning has been widely employed for various tasks related to document understanding and analysis. In this section, we review relevant studies that have utilized deep learning techniques for similar tasks, providing insights into their approaches and contributions.

- **"A robust system for document layout analysis using multilevel homogeneity structure" by Tran et al. (2017*) (Tran et al., 2017; A robust system for document layout analysis using multilevel homogeneity structure - ScienceDirect, 2023)*:**

This paper proposes a new method for document layout analysis. The method is based on the idea of a multilevel homogeneity structure, which is a way of representing the document image as a hierarchy of regions with similar properties. The method can identify different regions of interest in the document image, such as text, images, tables, and lists. The method is also able to identify the structure of the document, such as the page layout and the relationships between different regions.

- **"Analyzing Document Layout with LayoutParser" by Winastwan (2020*)*** *(Analyzing Document Layout with LayoutParser | by Ruben Winastwan | Towards Data Science, 2023)***:**

This article introduces LayoutParser, a Python library that offers pre-trained deep learning models for document layout analysis. LayoutParser provides an easy-to-implement solution for detecting the layout of document images from various sources. The library offers a range of functionalities, including text block detection, table extraction, and hierarchical layout analysis, making it a versatile tool for document layout analysis tasks.

- **"DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis" by Zhang et al. (2022)** *(Pfitzmann et al., 2022a; b; DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation | Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023)***:**

This paper introduces DocLayNet, a large human-annotated dataset specifically designed for document layout analysis. The dataset comprises over 100,000 document images with detailed ground-truth annotations for layout components such as text blocks, tables, figures, and captions. DocLayNet aims to provide a more challenging and diverse dataset compared to existing ones, enabling the training and evaluation of deep learning models for document layout analysis.

The above studies showcase the application of deep learning techniques in document layout analysis. The proposed deep learning-based system, along with the availability of pre-trained models and datasets like LayoutParser and DocLayNet, contribute to the advancement of document layout analysis tasks. These studies highlight the effectiveness and potential of deep learning approaches in accurately analyzing and understanding the layout structures within documents.

2.4 Evaluation of challenges and limitations in existing approaches

While deep learning methods have shown great promise in document layout analysis, there are still several challenges and limitations that researchers have encountered. In this section, we evaluate some of the common challenges and limitations observed in existing approaches.

- **Limited labeled data:**

One significant challenge in document layout analysis is the scarcity of large-scale labeled datasets. Annotated data for layout analysis tasks, such as text block classification, table detection, and graphical element recognition, is often time-consuming and costly to obtain. This limitation hampers the development of deep learning models that require substantial labeled data for effective training and generalization.

- **Generalization to diverse document layouts:**

Existing approaches often struggle to generalize well to diverse document layouts. Document layouts can vary significantly across domains, languages, and document types, posing a challenge for models trained on specific datasets. Deep learning models may not perform optimally on unseen layouts, requiring further adaptation or fine-tuning to ensure robust performance across diverse document collections.

- **Handling complex visual elements:**

Complex visual elements, such as overlapping text, non-rectangular shapes, and intricate graphics, pose challenges for existing deep-learning approaches. Models may struggle to accurately localize and understand these elements, leading to suboptimal performance in layout analysis tasks. Capturing fine-grained details and complex spatial relationships remains an ongoing challenge in deep learning-based document layout analysis.

- **Lack of interpretability:**

Deep learning models often lack interpretability, making it difficult to understand the decision-making process and the rationale behind layout analysis predictions. This limits the model's usability in scenarios where interpretability is crucial, such as legal document analysis or regulatory compliance. Developing explainable deep learning models for document layout analysis is an important area of research to enhance transparency and trust.

- **Computational requirements**

Deep learning models typically require significant computational resources, including high-performance GPUs and large amounts of memory, for training and inference. This poses challenges for researchers and practitioners with limited access to such resources, restricting the widespread adoption of deep learning approaches in document layout analysis. Optimizing models for efficiency without sacrificing performance is a critical consideration.

Addressing these challenges and limitations is crucial for advancing the field of document layout analysis. Researchers are actively working to create larger and more diverse datasets, develop models with improved generalization capabilities, handle complex visual elements, enhance interpretability, and optimize computational requirements. By addressing these limitations, future deep learning approaches can overcome existing challenges and provide more robust and practical solutions for end-to-end document layout analysis.

3. Data Preprocessing

3.1 Description of the DocBank dataset

The DocBank dataset serves as a benchmark dataset for document layout analysis and provides a comprehensive collection of documents for training and evaluation purposes. In this section, we provide a detailed description of the DocBank dataset, including its characteristics and composition.

- **Dataset Collection:**

The DocBank dataset is compiled from various sources, including academic papers, reports, news articles, and legal documents. It aims to encompass a wide range of document types and layouts, representing the diversity of real-world documents. The dataset collection process involves careful selection and curation to ensure a representative sample of document layouts.

- **Document Types and Layouts:**

The DocBank dataset consists of documents with diverse layouts, including but not limited to single-column and multi-column layouts, structured forms, tables, figures, captions, headings, and footnotes. This variety allows researchers to tackle different layout analysis tasks, such as text extraction, element classification, and semantic understanding.

- **Annotation Process**:

The documents in the DocBank dataset are manually annotated by experts to provide ground truth labels for layout analysis tasks. The annotation process involves labeling text regions, graphics, tables, headers, footers, and other layout elements present in the documents. This annotation provides a reference for evaluating the performance of document layout analysis models.

- **Annotation Guidelines:**

To ensure consistency and accuracy, annotation guidelines are established to guide the annotators in marking the document layouts. These guidelines define the criteria for labeling different layout elements, handling complex cases, and resolving ambiguities. The guidelines help maintain a high level of quality and facilitate reliable evaluations of layout analysis models.

- **Dataset Size and Statistics:**

The DocBank dataset is designed to be large-scale, comprising thousands of documents with varying lengths and complexities. The dataset statistics include information about the number of documents, the average number of pages per document, and the distribution of layout elements. These statistics provide insights into the characteristics and composition of the dataset.

The DocBank dataset serves as a valuable resource for training, validating, and benchmarking document layout analysis models. Its diverse collection of document types and layouts, along with manual annotations, enables researchers to develop and evaluate robust algorithms for end-to-end document layout analysis. The dataset's availability fosters advancements in the field and facilitates comparisons and collaborations among researchers.

3.2 Data cleaning and normalization techniques

Data cleaning and normalization are crucial preprocessing steps to ensure the quality and consistency of the data used for document layout analysis. In this section, we discuss various techniques employed to clean and normalize the DocBank dataset before using it for training or evaluation.

- **Text Cleaning:**

Textual data within the documents may contain noise, such as special characters, punctuation marks, or inconsistent formatting. Text cleaning techniques are applied to remove or replace these unwanted elements. Common approaches include removing non-alphanumeric characters, converting text to lowercase, and handling abbreviations or acronyms to improve text quality and consistency.

- **Noise Removal:**

Document images may contain noise, artifacts, or background patterns that can interfere with layout analysis tasks. Noise removal techniques, such as denoising filters or thresholding, are applied to eliminate unwanted elements and enhance the clarity of the document content. This process helps improve the accuracy of subsequent layout analysis algorithms.

- **Image Preprocessing:**

For documents with image components, preprocessing techniques are applied to enhance image quality and remove distortions. Image normalization methods, such as contrast adjustment, histogram equalization, or resizing, are used to standardize image properties across the dataset. These techniques improve the performance of image-based layout analysis tasks.

- **Layout Standardization:**

In some cases, it is beneficial to standardize the layout of documents within the dataset. This involves aligning the position, orientation, and scale of layout elements to a common reference. Standardization ensures consistent spatial relationships and simplifies subsequent analysis tasks, such as text extraction or graphics recognition.

- **Data Augmentation:**

Data augmentation techniques are often employed to increase the diversity and size of the dataset. Augmentation methods include rotation, scaling, cropping, flipping, or introducing random noise to the documents. These techniques help improve the model's generalization capabilities by exposing it to a wider range of variations in document layouts.

- **Metadata Extraction:**

In addition to cleaning and normalizing the textual and visual content, metadata extraction techniques may be applied to capture document-level information, such as titles, authors, dates, or document types. This metadata can be useful for organizing and categorizing the dataset, facilitating more targeted analysis or downstream applications.

By applying these data cleaning and normalization techniques, the DocBank dataset is prepared for training and evaluation. These preprocessing steps ensure data consistency, enhance the quality of the dataset, and improve the performance of document layout analysis models.

3.3 Splitting the dataset into training, validation, and test sets

After preprocessing the DocBank dataset, it is essential to split the data into separate subsets for training, validation, and testing. This section discusses the process of dividing the dataset and provides details on the purpose of each subset.

- **Training Set:**

The training set is used to train the document layout analysis model. It comprises a significant portion of the dataset and is utilized to optimize the model's parameters through iterative learning. A larger training set helps the model learn complex patterns and generalize well to unseen documents.

- **Validation Set:**

The validation set is used to fine-tune the model during training and to assess its performance on unseen data. It serves as a development set for hyperparameter tuning, model selection, and early stopping criteria. The validation set provides feedback on the model's performance and allows for adjustments to improve its generalization capabilities.

- **Test Set:**

The test set is used to evaluate the final performance of the trained model. It represents unseen data that the model has not been exposed to during training or validation. The test set provides an unbiased assessment of the model's ability to generalize and perform accurately on real-world document layouts.

- **Splitting Strategy:**

The dataset splitting strategy depends on various factors, including the dataset size, the desired ratio of training to validation to test data, and the nature of the documents. Common strategies include random splitting, stratified splitting, or time-based splitting for chronological document collections. It is crucial to ensure that each subset represents a diverse range of document types and layouts.

- **Data Balance:**

When splitting the dataset, it is important to consider class balance, especially if there are imbalanced layout elements or document types. Techniques such as stratified sampling can be employed to maintain a proportional representation of different layout classes in each subset. This ensures that the model is trained and evaluated on a balanced distribution of document layouts.

The division of the DocBank dataset into training, validation, and test sets allows for effective model development and evaluation. It enables the model to learn from a substantial amount of data, fine-tune its parameters, and assess its performance on unseen documents. Proper splitting ensures unbiased evaluation and provides reliable measures of the model's capabilities in document layout analysis tasks.

3.4 Preprocessing steps specific to document layout analysis

Document layout analysis involves understanding and extracting structured information from documents. In addition to general data cleaning and normalization techniques, specific preprocessing steps are employed to prepare the data for document layout analysis tasks. This section describes the preprocessing steps that are specific to document layout analysis.

- **Page Segmentation:**

Page segmentation refers to the process of dividing a document into individual pages. In the case of scanned or digitized documents, this step involves separating each physical page to enable independent analysis. Page segmentation can be achieved through techniques such as detecting page boundaries based on white spaces or using layout analysis algorithms to identify logical divisions within a document.

- **Optical Character Recognition (OCR):**

OCR is a critical step in document layout analysis, particularly for extracting text information. OCR algorithms are applied to convert scanned or image-based documents into machine-readable text. This involves detecting and recognizing characters or words in the document images, enabling subsequent analysis tasks such as text extraction, classification, or semantic understanding.

- **Text Extraction and Parsing:**

Text extraction aims to identify and extract textual content from documents, including headings, paragraphs, captions, and other text regions. Techniques such as optical character recognition, rule-based approaches, or machine learning algorithms are employed to extract and parse the text accurately. This step allows for subsequent analysis of the textual content and enables tasks such as keyword extraction, topic modeling, or sentiment analysis.

- **Layout Analysis:**

Layout analysis involves understanding the spatial arrangement and relationships of various layout elements within a document. Techniques such as connected component analysis, contour detection, or deep learning-based methods are utilized to identify and classify layout components, including text blocks, tables, graphics, headers, footers, and page numbers. Layout analysis provides structural information about the document and aids in further analysis tasks.

- **Document Normalization:**

Document normalization techniques aim to standardize the representation and format of documents. This involves aligning and normalizing the orientation, scale, or aspect ratio of document pages, removing artifacts or noise, and standardizing fonts, sizes, or styles. Document normalization ensures consistency and facilitates the comparison and analysis of documents with varying layouts and formats.

- **Meta-information Extraction:**

Meta-information extraction involves extracting document-level metadata such as titles, authors, dates, or document types. This information can provide additional context and assist in categorizing or organizing the documents. Techniques such as named entity recognition, pattern matching, or machine learning algorithms are employed to extract relevant metadata from the documents.

By applying these preprocessing steps specific to document layout analysis, the data is prepared to undergo layout analysis tasks effectively. These steps enable the extraction of structured information, such as text and layout elements, from documents and facilitate subsequent analysis and understanding of the document content.

4. Proposed Methodology

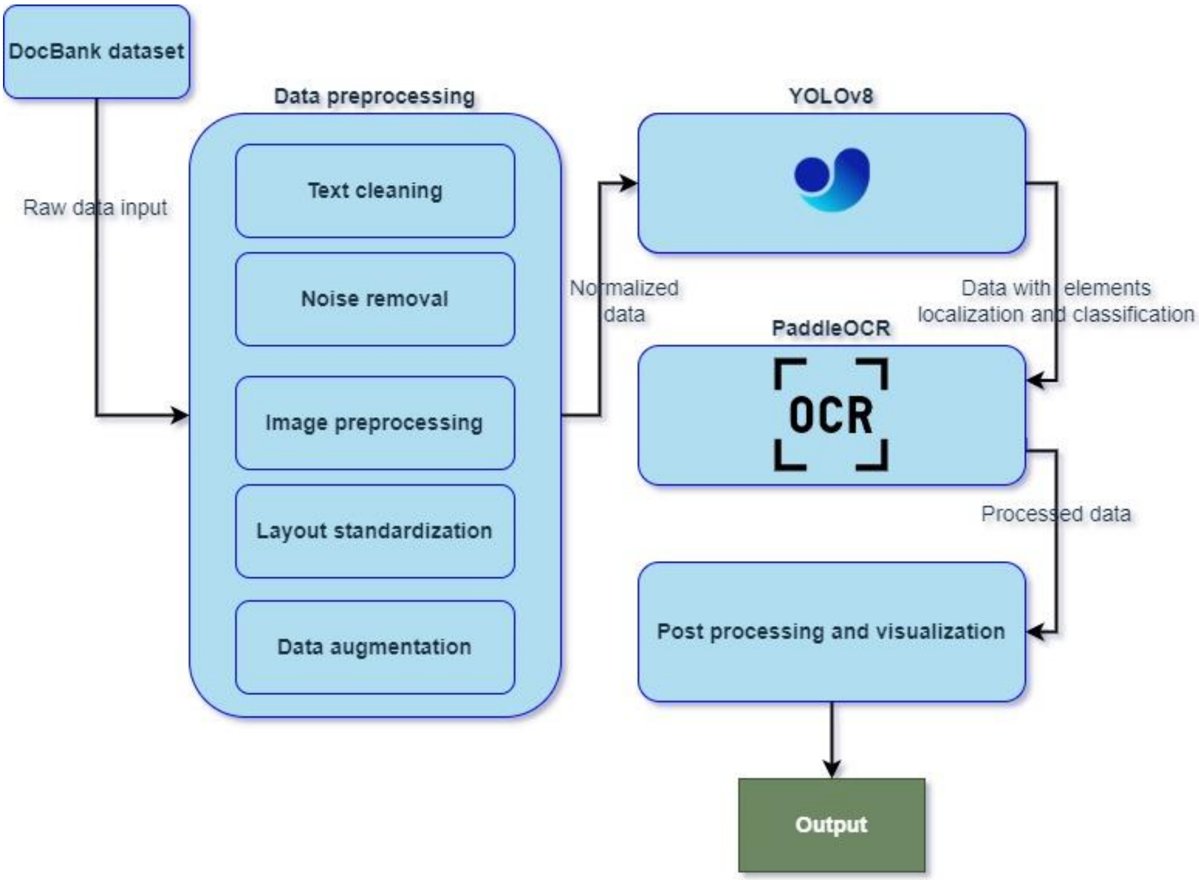


Figure 4.1 Proposed methodology

4.1 Detailed explanation of the end-to-end document layout analysis approach

The proposed methodology for end-to-end document layout analysis combines the YOLOv8 object detection model with the PaddleOCR optical character recognition (OCR) framework. This approach leverages the strengths of both models to achieve accurate and comprehensive document layout analysis. The following steps outline the methodology:

- **Preprocessing and Data Augmentation:**

The document images are preprocessed to enhance their quality and normalize the data. This may involve resizing, cropping, or applying image enhancement techniques to improve readability. Data augmentation techniques are also employed to increase the diversity and size of the dataset, allowing the model to learn from a broader range of document layouts.

- **Object Detection using YOLOv8:**

YOLOv8, a state-of-the-art object detection model, is used to identify and localize various layout elements in the document images. The YOLOv8 model is trained on a large-scale dataset, enabling it to detect text blocks, tables, graphics, headers, footers, and other relevant components. The model outputs bounding box coordinates and corresponding class labels for each detected layout element.

- **Optical Character Recognition (OCR) using PaddleOCR:**

The detected text regions from the YOLOv8 model are extracted and passed through the PaddleOCR framework for optical character recognition. PaddleOCR offers comprehensive text recognition capabilities and supports various languages and fonts. It recognizes the text within the bounding boxes, providing accurate and machine-readable versions of the textual content in the documents.

- **Layout Analysis and Semantic Understanding:**

The detected layout elements from YOLOv8 and the recognized text from PaddleOCR are combined to perform layout analysis and semantic understanding. This step involves analyzing the spatial relationships between the layout elements, identifying section headings, extracting tabular data, recognizing visual elements, and extracting relevant metadata such as titles, authors, and dates. Deep learning techniques, including recurrent neural networks (RNNs) or graph-based models, can be applied to model the layout structure and capture the contextual information.

- **Post-processing and Visualization:**

After the layout analysis and semantic understanding, post-processing steps are conducted to refine the results and improve accuracy. This may involve resolving conflicts between predicted layout elements, handling overlapping regions, or applying heuristics to enhance the interpretation of the layout structure. Finally, the analyzed layout information is visualized, providing a clear representation of the document's structure and facilitating further analysis or user interaction.

By combining the YOLOv8 object detection model with the PaddleOCR framework, the proposed approach achieves an end-to-end document layout analysis. It effectively identifies and localizes layout elements, recognizes text within those elements, and provides a comprehensive understanding of the document structure and content. This approach offers high accuracy and flexibility, making it suitable for various document layout analysis tasks.

4.2 Description of the deep learning architecture

The proposed end-to-end document layout analysis approach utilizes a deep learning architecture that combines the YOLOv8 object detection model and the PaddleOCR framework for optical character recognition. This section provides a more detailed description of the deep learning architecture employed in the methodology.

- **YOLOv8 Object Detection Model** *(Home - Ultralytics YOLOv8 Docs, 2023)*:

The YOLOv8 (You Only Look Once) model is a popular and efficient object detection architecture used for identifying and localizing objects within an image. It is based on a deep convolutional neural network (CNN) and achieves real-time object detection with high accuracy. YOLOv8 divides the input image into a grid and assigns bounding boxes to different objects present in the image. It simultaneously predicts the class labels and bounding box coordinates for each object. The model architecture consists of multiple convolutional layers, which extract hierarchical features from the input image, followed by fully connected layers for predicting the bounding boxes and class probabilities.

The YOLOv8 model is trained on a large-scale dataset, including annotated document layout images, to learn to detect various layout elements such as text blocks, tables, graphics, headers, footers, and more. The trained model is then used to identify and localize these layout elements within new document images.

- **PaddleOCR Framework** *(GitHub - PaddlePaddle/PaddleOCR: Awesome multilingual OCR toolkits based on PaddlePaddle (practical ultra lightweight OCR system, support 80+ languages recognition, provide data annotation and synthesis tools, support training and deployment among server, mobile, embedded and IoT devices), 2023)*

The PaddleOCR framework is utilized for optical character recognition (OCR) within the proposed methodology. PaddleOCR is an open-source OCR toolkit developed by PaddlePaddle, which provides a wide range of text recognition capabilities. The framework employs deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to recognize and extract text from images. It supports various languages, fonts, and text orientations, making it versatile for document layout analysis tasks. PaddleOCR is trained on large-scale OCR datasets, including annotated text images, to learn the visual patterns and linguistic structures of different characters and words. It then applies these learned patterns to recognize and extract text from the document images obtained from the YOLOv8 object detection step.

By combining the YOLOv8 object detection model and the PaddleOCR framework, the proposed deep learning architecture effectively detects and localizes layout elements within document images and recognizes the textual content present in those elements. This integrated architecture provides a robust and comprehensive solution for document layout analysis tasks, enabling accurate and automated understanding of document structures and content.

4.3 Training process and hyperparameter optimization

The training process and hyperparameter optimization play a crucial role in the proposed end-to-end document layout analysis approach. This section provides a more detailed explanation of the training process and the optimization of hyperparameters.

- **Training Process:**

The training process involves training the YOLOv8 object detection model on annotated datasets specific to document layout analysis. The following steps outline the training process:

- *YOLOv8 Training:*
  1. Dataset Preparation: Annotated datasets containing document images with labeled bounding boxes for layout elements are prepared. These datasets are split into training, validation, and test sets.
  2. Model Initialization: The YOLOv8 model is initialized with pre-trained weights on a large-scale object detection dataset, such as MS COCO, to capture general object detection features.
  3. Training: The YOLOv8 model is trained using the annotated document layout datasets. During training, the model optimizes its parameters through backpropagation and gradient descent to minimize the detection loss. Common loss functions used for object detection, such as the YOLO loss or the focal loss, can be employed.

- **Hyperparameter Optimization:**

Hyperparameter optimization is crucial to achieve the best performance of the deep learning models used in the proposed approach. Some key hyperparameters that can be optimized include:

- *Learning Rate*: The learning rate determines the step size during gradient descent and impacts the speed and convergence of the training process. It is typically set based on heuristics or through techniques like learning rate scheduling or adaptive optimization algorithms.
- *Batch Size:* The batch size determines the number of training samples processed together before updating the model's parameters. It affects the memory usage, computational efficiency, and generalization of the model.
- *Network Architecture*: The architecture of the deep learning models, such as the number and size of layers, can be customized to suit the document layout analysis task. Architectural choices may include adjusting the depth of the CNNs or the number of hidden units in the RNNs.

- Regularization Techniques: Regularization techniques, such as dropout or weight decay, can be applied to prevent overfitting and improve the model's generalization ability.
- Augmentation Strategies: Data augmentation techniques, such as random cropping, rotation, or brightness adjustments, can be employed to increase the diversity and robustness of the training data.

Hyperparameter optimization can be performed using techniques like grid search, random search, or more advanced methods such as Bayesian optimization or evolutionary algorithms. The goal is to find the hyperparameter values that result in the best performance on the validation set. By carefully training the YOLOv8 object detection model and optimizing the hyperparameters, the proposed approach aims to achieve high accuracy and robustness in document layout analysis tasks. The training process enables the models to learn the specific layouts.

4.4 Handling multi-class classification and bounding box regression

In the proposed end-to-end document layout analysis approach, multi-class classification and bounding box regression are essential components for accurately identifying and localizing different layout elements within document images. This section provides a more detailed explanation of how multi-class classification and bounding box regression are handled.

- Multi-Class Classification:

Multi-class classification is employed to assign class labels to the detected layout elements, such as text blocks, tables, graphics, headers, footers, and other relevant components. The following steps outline the process:

- *Dataset Annotation:* The training dataset is annotated with bounding boxes for each layout element and labeled with the corresponding class labels.
- *Training Objective*: The deep learning model, such as YOLOv8, is trained using a multi-class classification objective. The model aims to predict the class probabilities for each detected bounding box.
- *Loss Function:* The loss function used for multi-class classification is typically the categorical cross-entropy loss. It computes the error between the predicted class probabilities and the ground truth labels.
- *Prediction:* During inference, the model predicts the class probabilities for each detected bounding box. The class label with the highest probability is assigned to each layout element.

- Bounding Box Regression:

Bounding box regression is employed to accurately localize the layout elements within the document images by predicting the coordinates of their bounding boxes. The following steps outline the process:

- *Dataset Annotation:* The training dataset is annotated with bounding boxes representing the exact spatial extent of each layout element.
- *Training Objective*: The deep learning model, such as YOLOv8, is trained using a bounding box regression objective. The model aims to predict the coordinates of the bounding box (e.g., top-left corner coordinates, width, height) for each detected layout element.
- *Loss Function*: The loss function used for bounding box regression is typically a combination of localization and size-related terms, such as the mean squared error (MSE) loss or the smooth L1 loss. It computes the error between the predicted bounding box coordinates and the ground truth coordinates.
- *Prediction:* During inference, the model predicts the coordinates of the bounding box for each detected layout element, allowing for precise localization of the elements within the document images.

By incorporating multi-class classification and bounding box regression techniques, the proposed methodology enables accurate identification and localization of various layout elements in document images. The multi-class classification assigns appropriate class labels to each element, while the bounding box regression predicts the precise spatial coordinates of the elements. Together, these techniques provide a comprehensive solution for document layout analysis, facilitating further understanding and processing of the document structure and content.

4.5 Integration of post-processing techniques

Post-processing techniques are crucial in refining the results of the document layout analysis process and improving the overall accuracy and usability of the proposed approach. This section provides a more detailed explanation of the integration of post-processing techniques within the methodology.

- Text Refinement:

After performing optical character recognition (OCR) using the PaddleOCR framework, the recognized text may contain errors or inaccuracies. Post-processing techniques can be applied to refine the extracted text and improve its quality. Some common text refinement techniques include:

- *Spell Checking:* Spell-checking algorithms can be employed to identify and correct spelling errors in the recognized text. These algorithms utilize dictionaries or language models to compare the extracted text with a set of valid words and suggest corrections for misspelled words.

  - *Language Model Post-Processing:* Language models, such as recurrent neural networks (RNNs) or transformer models, can be used to improve the accuracy and coherence of the recognized text. These models leverage context and language patterns to correct grammatical errors and enhance the overall quality of the extracted text.
  - *Post-OCR Correction:* Techniques like rule-based or pattern-based correction can be employed to fix specific types of errors commonly encountered during OCR, such as misinterpreted characters, incorrect spacing, or noise removal.

- Layout Refinement:

The initial layout analysis results obtained from the YOLOv8 object detection model can be further refined to enhance the accuracy of the detected layout elements. Post-processing techniques for layout refinement may include:

  - *Geometric Constraints:* Geometric constraints can be applied to adjust the positions and orientations of the detected layout elements. For example, if a header is expected to appear at the top of a document page, geometric constraints can be used to enforce this position and correct any misalignments.
  - *Structural Analysis:* Structural analysis techniques can be employed to validate the relationships between different layout elements. For instance, the hierarchical structure of a document can be analyzed to ensure that captions are associated with the correct figures or tables.
  - *Noise Removal:* Post-processing steps can be implemented to eliminate false positive detections or remove noise in the layout analysis results. This can involve filtering out small or insignificant layout elements that are unlikely to be valid components of the document structure.

- Document Reconstruction:

Document reconstruction techniques can be applied to combine the refined layout elements and recognized text into a coherent representation of the original document. These techniques involve arranging the elements in the correct order and formatting them according to the document's structure. This can include tasks such as reconstructing paragraphs, aligning columns, or merging separate text blocks into cohesive sections.

By integrating post-processing techniques, the proposed methodology enhances the accuracy and quality of the document layout analysis results. The refinement of recognized text and layout elements improves the overall usability and reliability of the extracted information, enabling more effective downstream document processing and analysis tasks.

5. Experimental Setup

5.1 Resource requirements

The experimental setup involves the utilization of specific hardware and software to train and evaluate the proposed end-to-end document layout analysis approach. This section provides a more detailed description of the hardware and software components used in the experiments.

- **Hardware:**

The hardware used for training and evaluation typically includes a high-performance computing system with sufficient computational resources to handle the deep learning tasks involved in the proposed approach. The specific hardware configuration may vary depending on the scale and complexity of the document layout analysis tasks. Some common hardware components used are:

  - *Central Processing Unit (CPU):* A powerful CPU is essential for handling complex computations and data preprocessing tasks. CPUs with multiple cores or high clock speeds are beneficial for accelerating the training and evaluation processes.
  - *Graphics Processing Unit (GPU):* GPUs play a crucial role in accelerating the training and inference of deep learning models. GPUs are highly parallelized and efficient for matrix operations, making them ideal for the computationally intensive tasks involved in deep learning.
  - *Memory (RAM):* Sufficient memory capacity is required to accommodate the large datasets and model parameters used in document layout analysis. Higher RAM capacity allows for faster data access and better performance during training and evaluation.
  - Storage: Adequate storage capacity is necessary to store the datasets, pre-trained models, and checkpoints generated during the training process. Solid-state drives (SSDs) or high-capacity hard disk drives (HDDs) are commonly used for efficient data storage.

- **Software:**

The software stack used in the experimental setup includes various libraries, frameworks, and tools that facilitate the implementation and execution of the proposed document layout analysis approach. The specific software components utilized may include:

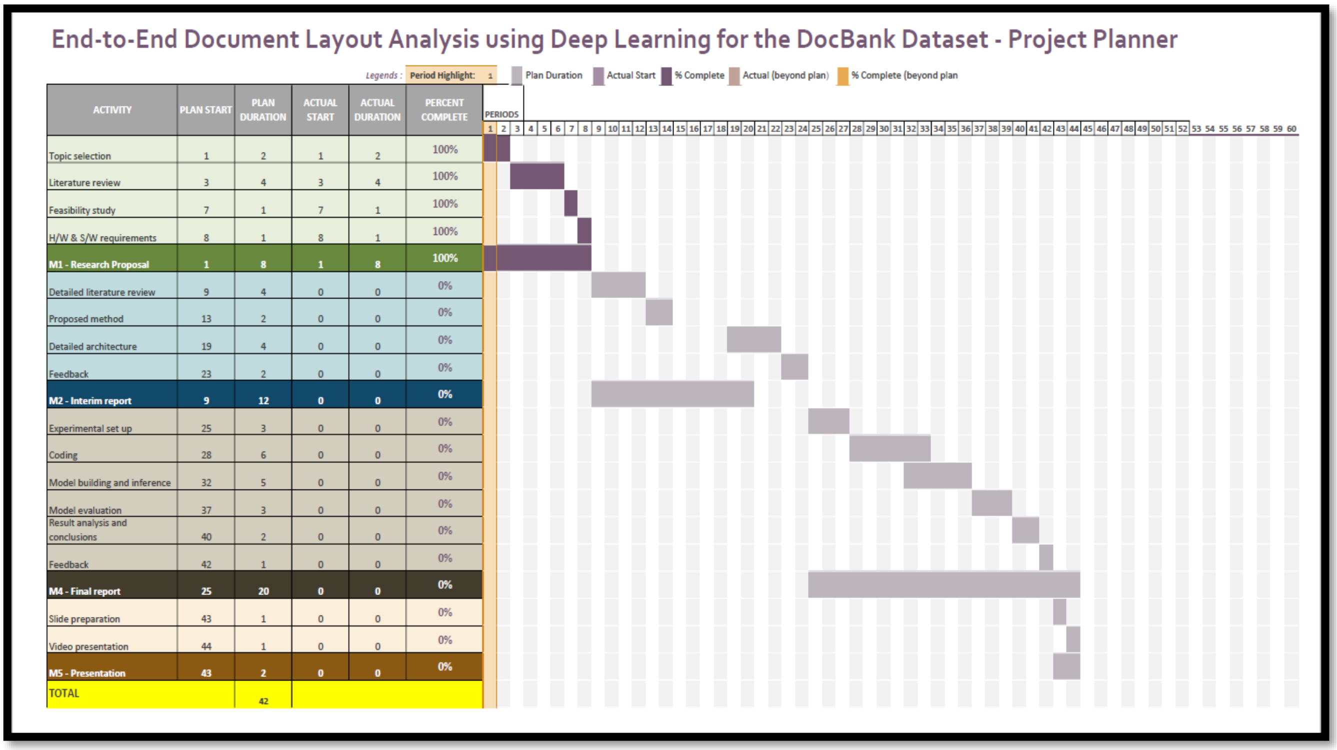
  - *Deep Learning Frameworks:* Deep learning frameworks such as TensorFlow, PyTorch, or Keras are typically employed to build and train the deep learning models. These frameworks provide a high-level interface for implementing complex neural network architectures and optimizing model parameters.
  - *Object Detection Framework:* The YOLOv8 model, as mentioned in the proposed approach, can be implemented using frameworks like Darknet, PyTorch-YOLO, or TensorFlow-YOLO. These frameworks provide pre-implemented versions of the YOLO architecture and associated training utilities.
  - *Optical Character Recognition (OCR) Framework:* The PaddleOCR framework, mentioned for text extraction, can be utilized to perform OCR tasks. PaddleOCR offers pre-trained models, APIs, and utilities for text extraction from document images.

- *Programming Languages:* The experimental setup may involve programming languages such as Python, which provides extensive libraries for deep learning, image processing, and data manipulation. Python's ecosystem enables efficient implementation of the proposed methodology.
- *Data Manipulation and Visualization:* Libraries like NumPy, Pandas, and Matplotlib are often employed for data manipulation, analysis, and visualization tasks. These libraries facilitate dataset preparation, data augmentation, and result visualization.
- Operating System: The experiments can be conducted on popular operating systems such as Linux, Windows, or macOS, depending on the compatibility of the selected software components.

By utilizing appropriate hardware components and software tools, the experimental setup ensures efficient training, evaluation, and implementation of the proposed end-to-end document layout analysis approach. The hardware's computational power, along with the software's capabilities, enables effective deep learning model training and evaluation, leading to accurate and robust results.

6. Research plan

6.1 Gantt chart



End-to-End Document Layout Analysis using Deep Learning for the DocBank Dataset - Project Planner

Legends : Period Highlight: 1 ▨ Plan Duration ▨ Actual Start ■ % Complete ▨ Actual (beyond plan) ▨ % Complete (beyond plan

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Topic selection	1	2	1	2	100%
Literature review	3	4	3	4	100%
Feasibility study	7	1	7	1	100%
H/W & S/W requirements	8	1	8	1	100%
M1 - Research Proposal	**1**	**8**	**1**	**8**	**100%**
Detailed literature review	9	4	0	0	0%
Proposed method	13	2	0	0	0%
Detailed architecture	19	4	0	0	0%
Feedback	23	2	0	0	0%
M2 - Interim report	**9**	**12**	**0**	**0**	**0%**
Experimental set up	25	3	0	0	0%
Coding	28	6	0	0	0%
Model building and inference	32	5	0	0	0%
Model evaluation	37	3	0	0	0%
Result analysis and conclusions	40	2	0	0	0%
Feedback	42	1	0	0	0%
M4 - Final report	**25**	**20**	**0**	**0**	**0%**
Slide preparation	43	1	0	0	0%
Video presentation	44	1	0	0	0%
M5 - Presentation	**43**	**2**	**0**	**0**	**0%**
TOTAL		42			

6.2 Risk mitigation and contingency plan

As this research work has been done in parallel with the professional commitment, there are major challenges on meeting the timelines. The risks associated with the research work are mentioned in the below table along with the contingency plan.

Table 6.2.1 Risk and contingency

Risks	Contingency
Professional commitments to supersede the research commitments and priorities or some leaves	• Plan for the buffer time in project plan. • Identify the potential options of extensions with the help of University/UpGrad administrations.
Resources unavailability or technical impediments	• Identify the best cloud high performance resources for model evaluation. • Have frequent connects with supervisor and plan proactive steps to avoid technical impediments.

References

Anon (2023) *[2012.14740] LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding*. [online] Available at: https://arxiv.org/abs/2012.14740 [Accessed 14 Jul. 2023].

Anon (2023) *[2104.02874] Document Layout Analysis via Dynamic Residual Feature Fusion*. [online] Available at: https://arxiv.org/abs/2104.02874 [Accessed 14 Jul. 2023].

Anon (2023) *A novel deep learning method based on attention mechanism for bearing remaining useful life prediction - ScienceDirect*. [online] Available at: https://www.sciencedirect.com/science/article/abs/pii/S1568494619307008 [Accessed 14 Jul. 2023].

Anon (2023) *A robust system for document layout analysis using multilevel homogeneity structure - ScienceDirect*. [online] Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417417303469 [Accessed 14 Jul. 2023].

Anon (2023) *Analyzing Document Layout with LayoutParser | by Ruben Winastwan | Towards Data Science*. [online] Available at: https://towardsdatascience.com/analyzing-document-layout-with-layoutparser-ed24d85f1d44 [Accessed 14 Jul. 2023].

Anon (2023) *BINYAS: a complex document layout analysis system | SpringerLink*. [online] Available at: https://link.springer.com/article/10.1007/s11042-020-09832-3 [Accessed 14 Jul. 2023].

Anon (2023) *Convolutional Neural Network: An Overview*. [online] Available at: https://www.analyticsvidhya.com/blog/2022/01/convolutional-neural-network-an-overview/ [Accessed 14 Jul. 2023].

Anon (2023) *DiT: Self-supervised Pre-training for Document Image Transformer | Proceedings of the 30th ACM International Conference on Multimedia*. [online] Available at: https://dl.acm.org/doi/abs/10.1145/3503161.3547911 [Accessed 13 Jul. 2023].

Anon (2023) *DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation | Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. [online] Available at: https://dl.acm.org/doi/abs/10.1145/3534678.3539043 [Accessed 14 Jul. 2023].

Anon (2023) *Document Layout Analysis via Dynamic Residual Feature Fusion | IEEE Conference Publication | IEEE Xplore*. [online] Available at: https://ieeexplore.ieee.org/abstract/document/9428465 [Accessed 14 Jul. 2023].

Anon (2023) *GitHub - PaddlePaddle/PaddleOCR: Awesome multilingual OCR toolkits based on PaddlePaddle (practical ultra lightweight OCR system, support 80+ languages recognition, provide data annotation and synthesis tools, support training and deployment among server, mobile, embedded and IoT devices)*. [online] Available at: https://github.com/PaddlePaddle/PaddleOCR [Accessed 14 Jul. 2023].

Anon (2023) *Home - Ultralytics YOLOv8 Docs*. [online] Available at: https://docs.ultralytics.com/ [Accessed 14 Jul. 2023].

Anon (2023) *Recurrent Neural Network. So Why do we need a recurrent neural… | by Devanshi | DataDrivenInvestor*. [online] Available at: https://medium.datadriveninvestor.com/recurrent-neural-network-58484977c445 [Accessed 14 Jul. 2023].

Anon (2023) *Transformer Encoder-Decoder architecture, taken from Vaswani et al. [9]... | Download Scientific Diagram*. [online] Available at: https://www.researchgate.net/figure/Transformer-Encoder-Decoder-architecture-taken-from-Vaswani-et-al-9-for-illustration_fig2_338223294 [Accessed 14 Jul. 2023].

Bhowmik, S., Kundu, S. and Sarkar, R., (2021) BINYAS: a complex document layout analysis system. *Multimedia Tools and Applications*, [online] 806, pp.8471–8504. Available at: https://link.springer.com/article/10.1007/s11042-020-09832-3 [Accessed 13 Jul. 2023].

Chen, Y., Peng, G., Zhu, Z. and Li, S., (2020) A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Applied Soft Computing*, 86, p.105919.

Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C. and Wei, F., (2022) DiT: Self-supervised Pre-training for Document Image Transformer. *MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia*, [online] pp.3530–3539. Available at: https://dl.acm.org/doi/10.1145/3503161.3547911 [Accessed 13 Jul. 2023].

Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z. and Zhou, M., (n.d.) *DocBank: A Benchmark Dataset for Document Layout Analysis*. [online] Online. Available at: https://github.com/doc-analysis/DocBank.

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S. and Staar, P., (2022a) DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. [online] Available at: https://doi.org/10.1145/3534678.3539043 [Accessed 14 Jul. 2023].

Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S. and Staar, P., (2022b) DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, [online] pp.3743–3751. Available at: https://dl.acm.org/doi/10.1145/3534678.3539043 [Accessed 14 Jul. 2023].

Tran, T.A., Oh, K., Na, I.S., Lee, G.S., Yang, H.J. and Kim, S.H., (2017) A robust system for document layout analysis using multilevel homogeneity structure. *Expert Systems with Applications*, 85, pp.99–113.

Wu, X., Hu, Z., Du, X., Yang, J. and He, L., (2021a) DOCUMENT LAYOUT ANALYSIS VIA DYNAMIC RESIDUAL FEATURE FUSION. *Proceedings - IEEE International Conference on Multimedia and Expo*.

Wu, X., Hu, Z., Du, X., Yang, J. and He, L., (2021b) Document Layout Analysis via Dynamic Residual Feature Fusion. *Proceedings - IEEE International Conference on Multimedia and Expo*. [online] Available at: http://arxiv.org/abs/2104.02874 [Accessed 14 Jul. 2023].

Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M. and Zhou, L., (2020a) LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. [online] Available at: http://arxiv.org/abs/2012.14740 [Accessed 14 Jul. 2023].

Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M. and Zhou, L., (2020b) LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, [online] pp.2579–2591. Available at: https://arxiv.org/abs/2012.14740v4 [Accessed 14 Jul. 2023].