Document Sanctity Check Using Machine Learning

Thesis by

Muhammed Faris Mukthar M V

223037

In Partial Fulfillment of the Requirements for the

Degree of

M.Sc. Computer Science With Specialization in Data Analytics

SCHOOL OF DIGITAL SCIENCES

KERALA UNIVERSITY OF DIGITAL SCIENCE, INNOVATION AND TECHNOLOGY

Trivandrum, Kerala

Supervisor: Dr. Anoop V. S.

2024

BONAFIDE CERTIFICATE

This is to certify that the project report entitled "Document Sanctity Check Using Machine Learning" submitted by **Muhammed Faris Mukthar M V (Reg. No: 223037)** in partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics is a bonafide record of the work carried out at School of Digital Sciences under our supervision.

Guide	**Internal Guide**
Dr. Anoop V. S.	Dr. Anoop V. S.
Research Officer,	Research Officer,
School of Digital Sciences	School of Digital Sciences

DECLARATIONS

I, Muhammed Faris Mukthar M V, student of Masters of Science in Computer Science with specialization in Data Analytics, hereby declare that this report is substantially the result of my own work, except where explicitly indicated in the text, and has been carried out during the period February 2024 to June 2024.

Place: Thiruvananthapuram

Date: July - 2024

ACKNOWLEDGEMENTS

I would like to offer my sincere gratitude and appreciation to everyone who has helped and encouraged me along the way. I could not have accomplished my goals and gotten to where I am now without their support, direction, and encouragement..

I want to start by gratefully thanking **Dr. Anoop V. S.**, my guide, for his continuous support and guidance during my research endeavour. His knowledge, counsel, and unwavering encouragement have been extremely helpful in directing my work and inspiring me to pursue greatness.

For providing excellent instruction and a supportive study atmosphere, **Prof. John Eric Stephen** and our institution, the institution of Digital Sciences, have my sincere gratitude. I am really grateful to **Dr. Saji Gopinath**, our Vice Chancellor, for providing me with the opportunity to grow in such a remarkable university as Digital University Kerala. Their commitment to promoting an intellectually curious society has had a major influence on both my academic and personal development.

I would be negligent if I did not acknowledge the help and comprehension of my friends and family. Throughout my journey, their steadfast support, words of wisdom, and trust in me have been a continual source of strength. I consider myself very lucky to have such a strong support network.

ABSTRACT

Technological advancements have created new opportunities for document analysis tasks, such as resume shortlisting, invoice extraction, and bill summarizing, by extracting content efficiently. This project focuses on verifying documents by assessing their integrity, authenticity, and validity using machine learning techniques. Our aim was to develop an automated machine learning model for ensuring document sanctity. To achieve this, we utilized transformers and detection models. Initially, we collected various documents relevant to our problem, including those with headers, those without headers, and various structured forms containing seals and signatures. Extracting meaningful information from these documents required a model capable of understanding their structure and context. For this purpose, we employed LayoutLMv3, a state-of-the-art model that can be fine-tuned for such tasks. Through extensive research, we identified LayoutLMv3 as the most suitable model for our needs. Additionally, we used Faster R-CNN as our detection model.We developed a fine-tuned LayoutLMv3 model for classifying documents based on their structure and content. Integrating the Faster R-CNN model enabled us to detect signs and seals within documents, thereby verifying their integrity. By combining these two models, we created a seamless pipeline that automates the document sanctity check process. This innovative approach is expected to streamline various processes and reduce bottlenecks significantly.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

With innovation highly taking over most aspects of modern life, a time undoubtedly comes when automated document verification will have to be apparent in all industries. The most prevalent form of document verification[28] at the moment is manual verification. This process relies exclusively on human inspection to ascertain whether a document is legitimate and original. This is a labor- and time-consuming process, resulting in processing delays and inefficiencies attributed to the enormous workload. Additionally, human error is so common that the precision and reliability of their output vary. Thus, with the need for effective, accurate, and reliable verification systems, this has catalyzed computer vision and machine learning technologies. Currently, every organization is mainly run by digital documents and procedures in every vital operation; therefore, the verification becomes very crucial with maximum speed and accuracy.

The overall objective of the project is to come up with a design and simple working system that could reasonably determine, detect, and classify certificates for their type and validate them. Rejecting a document not needed in cases of uncertainty, those instances shall be marked to be detected by the machine for manual verification to ensure ambiguities are checked.

This work contains many important sections as follows. First, generating the dataset ensures diversity in representation and structural arrangement of various documents[12]. This is because it obtains a varied set of documents containing photos from the customer. Among the several kinds of documents falling into this main category are titled, untitled, and form-based papers. This forms the first step of the process and hence an important one since it offers an extensive dataset that captures different forms, styles, and complexity found in real-world situations. These documents will then be annotated using Label Studio to provide the ground truth data necessary for training and assessing the detection model through sign-seal detection annotation, key-value pair annotation for structural analysis of the text[14].

One of the core elements within our methodology is the LayoutLMv3 model. We are going to utilize this specialized model that was created to understand and treat document layouts complicated in nature by use of text, layout, and visual features.

This is very useful in document processing tasks such as form interpretation, extraction of tables, and document categorization since it improves the earlier versions of LayoutLM in handling multimodal data[4]. It tokenizes the text using layout embeddings that help to capture spatial arrangements and further provide context with visual features from a pre-trained CNN[6]. The model learns how to link tokens by their placements and content using its transformer encoder with self-attention processes. This will be able to enable LayoutLMv3 to identify entities and hence assign keys to the corresponding values.[1]

We use Faster R-CNN, a model specifically designed to understand and deal with complicated object identification tasks through its integration of region proposal networks with convolutional neural networks for detection. It has improved in region-based detection a lot compared to former versions, and works efficiently for all tasks concerning object detection, instance segmentation, and image classification. It applies a region proposal network for candidate generation of possible object bounding boxes and leverages a previously trained CNN to extract information from these regions. Through RoI pooling and its fully connected layers, it learns the relationships between regions with respect to contextual and spatial information. Thus, it would be able to accurately locate objects and then dimensionally relate the boxes to their corresponding class labels.[7]

It means that such technologies are in place to drastically cut the workload that was required by manual inspection, complemented by document classification and verification. This is highly useful automation in high-volume settings where human inspectors are prone to error and fatigue. In this way, organizations can efficiently and effectively simplify these processes in order to boost operational reliability and productivity. The benefits of such a system extend beyond the realm of solely operational efficiency. There are also some broader implications regarding security and compliance when one considers the automated verification of digital certificates. Verification of certificates bears a prime role in upholding trust and integrity, especially in sectors like business, government, and education. Automated solutions not only reduce the risk of fraud but can also provide true document security on every count, besides offering the same result always. These technologies help organizations enhance the reliability of documents and fight off fraud.

The project mainly aims at high accuracy and dependability of certificate verification. Besides, high-accuracy techniques of cutting-edge technology improve accuracy, efficiency, and make things much simpler and quicker. Certificate detection

and verification automation generally help organizations reap major benefits in terms of productivity and operational dependability. Ultimately, this is an all-inclusive approach towards controlling the challenges of manual document verification through state-of-the-art technologies in the most efficient domains and techniques. We, in regard to this, prepared and checked methodologies for dataset generation, execution, model running, and model testing through test files to reach the optimal outputs. This project progress effort targets systems that will automate document verification.

In summary, our results provide evidence on how far our approach is working toward extracting data from specific kinds of documents. We were able to precisely parse and extract useful data using our algorithm from a test set of documents, and on the trained dataset, we also obtained excellent results for Faster R-CNN[22] applied to seal and signature detection. These results show how the mentioned methods can be successfully applied and change the way organizations validate documents, guaranteeing more accuracy, efficiency, and security. Hence, such an introduction provides an in-depth explanation of the methodologies, findings, and implications of the project and solely tracks the overview of our approach to enhancing document verification through automation.[26]

1.1 Problem statement

The need for automated document verification is growing clearer by the day in every sector relating to this new digital age. Even to this very day, the more predominant type of document verification technique has been the manual one, which relies solely on human eyes to check and verify the authenticity and integrity of the documents at hand. Further, this methodology is labor- and time-intensive, hence processing-deficient and slow due to a huge amount of work. Moreover, human error is rampant, which reduces consistency in accuracy and reliability. Such reliance on manual verification is untenable in an age where every organization relies on digital documents and workflows for core operations. Accordingly, the need for fast, accurate, and reliable verification systems also becomes more compelling because of innovation pressures from improved computer vision and machine learning technologies.

1.2 Objectives

This project's main objective is to demonstrate and develop a simple system that can reliably identify, classify, and validate certificates according to their type, and reject unnecessary documents. Specific objectives include:

- Automating the classification and verification of documents to reduce the workload of manual inspection.
- Enhancing operational efficiency and reliability by minimizing human errors and processing delays.
- Utilizing advanced models like LayoutLMv3 for understanding structured documents and Faster R-CNN for detecting seals and signatures.
- Ensuring that any uncertainties detected by the system are marked for manual verification to thoroughly examine ambiguities.
- Demonstrating the practical implementation and effectiveness of these technologies in high-volume scenarios where human inspectors are prone to errors and fatigue.

1.3 Organization of the report

This structure of the report will, therefore, enable us to provide a comprehensive and very coherent narrative that leads the reader through the project, from problem statement to final conclusions and recommendations. Under this approach, every aspect of the project will be rigidly explored and expressed clearly to enable greater understanding of the work undertaken and its implications. The methodology of this project, its findings, and the eventual implications in enhancing document verification by automation are well arranged in this report. Chapter I: Introduction presents the problem statement and the objectives of this project. A review of the available literature to provide the due background information relevant to automated document verification is presented in Chapter II: Background and Related Studies. Chapter III: Materials and Methods elaborates on various tools and models used in the project, such as LayoutLMv3, Faster R-CNN, and Label Studio. The roles played and functionalities of the same are explained. Chapter IV: Proposed Approach describes the methodology adopted for this project and further contains a comparison between the proposed technique and other existing methods for proposing advantages. Chapter V: Results and Discussions presents the results obtained from the experiments and gives an in-depth discussion of the findings that prove the efficiency of the proposed approach. Finally, Chapter VI: Conclusion and Future Work summarizes key outcomes of the project, pointing to future directions of the work, therefore leading to further research and development for the enhancement of automated document verification systems.

Chapter 2

BACKGROUND AND RELATED STUDIES

In the last couple of years, document understanding field, especially VDU and RE in the domain of NLP, has seen remarkable developments. These developments have changed the way research[1] in document analysis tasks is approached . Typically, every relation extraction task used to be limited only in a sentence; the transition to document-level tasks spurred a need for models that could integrate textual and visual information easily. This has been of particular importance in dealing with visually rich documents, where spatial arrangement and pictorial data take the lead in content understanding and extraction.

Firstly, the leading models that came into being to handle the intricacies brought on board by VRDs are LayoutLM, LayoutLMv2, LayoutLMv3,[20] and BROS. These models resort to methods such as positional encoding and further integrate absolute and relative spatial information to retrieve meaningful relations between elements in documents. More specifically, LayoutLMv3 revealed quite amazing results in terms of increasing the performance boost for the automated document verification system, although often complemented by other technologies like Faster R-CNN for entire document analysis. Such innovations make verification processes easier and have wide-ranging potential uses in many other industries, underscoring the fact that multi-modal approaches to document understanding and processing are assuming an increasingly important role.

The SC-Faster R-CNN[26] algorithm represents a radical improvement in object detection performance, particularly where the objects are obscured or distorted. SC-Faster R-CNN incorporates skip pooling and the fusion of contextual information in its network to improve object detection accuracy, although the challenges presented by deformed, rotating, and camouflage objects remain. Future directions include enhancing detection efficiency for such complex scenarios and optimizing real-time system performance so that an object detection system can be effectively applied in quite a few diversified real-world applications.

LayoutLMv3, illustrated in Layoutlmv3: Pre-training for document ai with unified text and image masking [12], is quite a ground-breaking multimodal pre-trained model designed specifically for Document AI applications. Contrasted with tra-

ditional methods applied separately by CNNs for image embeddings, it combines text, image, and multimodal representations within a single transformer architecture. This does not bring only ease in terms of model parameters; it also eliminates region annotations, hence bringing efficiency during deployment. LayoutLMv3[18] excels across a wide variety of tasks, including form understanding, receipt analysis, document image classification, and layout analysis, by exploiting a powerful pre-training framework that involves Masked Language Modeling, Masked Image Modeling, and Word-Patch Alignment. From text-centric to image-centric tasks, this flexibility makes it quite versatile for Document AI applications, ensuring reliability and simplicity for real-world document comprehension and analysis.

The very newest research in Fast R-CNN[7], introduced by Girshick in 2015, puts a strong point on how this model is efficient with improvements over other existing models dealing with object detection, such as R-CNN and SPPnet. Experimental evaluations firmly put Fast R-CNN at the forefront because of its high degree of improvement in detection speed, improving the speed by 10-100 times over previous methods. The success behind this approach lies within the feature extraction methodology, which is based on max-pooling over feature maps by means of spatial pyramid pooling techniques. This has drastically reduced training time while further improving on detention capabilities. Much effort in object detection currently goes into the refinement of these methodologies so that close-to-equality performance between sparse and dense object detection scenarios can be achieved. As such, this accelerates the overall object detection processes in practical applications.

Advanced techniques, such as those explored in the Kaggle notebook "Fine-Tune LayoutLM on SROIE Dataset," further raise the bar in this landscape of document processing. This paper presents a step-by-step guide on how to fine-tune LayoutLM on target domains with the SROIE dataset[15]to enhance its performance over receipt and invoice documents. This notebook, starting with the introduction of LayoutLM, a Transformer-based model good at text and layout information integration, goes ahead to instruct the key steps in dataset preparation with respect to data cleaning, annotation, and formatting according to the input requirements of the model. Core to its focusing is the process of fine-tuning pipeline configuration, hyperparameters tuning, and metrics optimization tailored for receipt OCR tasks. Evaluation methodologies outlined in the notebook underline rigorous validation and accuracy assessment on dedicated datasets, underscoring the efficacy of LayoutLM to automate document processing workflows for receipts and invoices.[2]

In the Google Research blog post, "Extracting Structured Data from Templatic Documents," [19]it introduces new methodologies that extract structured data from templatic documents with quite consistent layouts but variable content. It calls for leveraging techniques in machine learning and computer vision to bring about accuracy and efficiency in the challenges inherent in automated data extraction from such documents. Some remarkable progress in regard to OCR and deep learning models for parsing structured information like dates, amounts, and names from invoices and forms are discussed. These developments will be able to automate document workflows across wide-ranging industries reliant on structured data extraction, thus fostering operational efficiency and data accuracy.

The other paper, "Synergizing Optical Character Recognition: Comparative Analysis and Integration of Tesseract, Keras, Paddle, and Azure OCR"[16], contains a comparative analysis helpful in optimizing the workflows of OCR. On this matter, this research reviews well-known OCR technologies such as Tesseract, Keras OCR, PaddleOCR, and Azure OCR by benchmarking them on parameters like accuracy, speed, and integration ease. Through the analysis, one can find out that every OCR tool has strengths and weaknesses, making each of them appropriate in different cases and environments. This also exploits strategies on how multiple OCR engines can be synergistically integrated to enhance the overall performance and reliability for any application in OCR, answering the diverse requirements of applications and operational constraints.

The chapter "Evaluation of Generic Deep Learning Building Blocks for Segmentation of Nineteenth Century Documents" on IntechOpen[21] shares an in-depth review on deep learning applied to historical document segmentation. In the mentioned research, the efficiency of convolutional and recurrent neural networks is evaluated against common problems of nineteenth-century documents: character fonts, degradation of paper quality, and complex layouts. Experimental findings underline the potential of deep learning methodologies for document segmentation to transform state-of-art capabilities related to text-image extraction in historical documents. This chapter acts as a scholarly reference as to how modern AI applications are being used today for the preservation and processing of historical documents.

Thirdly, new developments associated with object detection, particularly "Unbiased Faster R-CNN for Single-Source Domain Generalized Object Detection"[17] at CVPR 2024, were about the challenges of single-source domain generalization for object detection. Taking into account biases caused by unseen domains, it proposes

an Unbiased Faster R-CNN framework that integrates structural causal models with causal attention mechanisms. These innovations enhance both the robustness and generalizability of models by mitigating biases due to scene and object attribute confounders and demonstrate improvements in performance across a wide range of scenarios. Accordingly, the results clearly spell out that the leading role, with regard to enhancing flexibility toward real-world challenges of the current state in object detection systems, was played by adversarial training methods and further thematic steps in the areas of domain-generalized object detection capabilities.

Besides, "Multi-adversarial Faster-RCNN with Paradigm Teacher for Unrestricted Object Detection" has investigated new methodologies in object detection by proposing the Multi-adversarial Faster-RCNN framework. MAF is designed to cope with certain challenges in unrestricted object detection scenarios that have very different representative appearances of objects and backgrounds; it holds a mechanism named paradigm teacher to enhance model robustness and performance. Experimental validation proves that MAF outperforms existing Faster-RCNN models and performs broadly on benchmark datasets. The paper generalizes across challenging and very diverse environments in promoting strategies of adversarial training as part of the development that will foster object detection systems[9].

In the paper "Information Extraction from Financial Statements based on Visually Rich Document Models," at the Encontro Nacional de Inteligência Artificial e Computacional [3], it provides information on how Visually Rich Document Models are to realize and utilize robust information extraction from financial documents. VRDMs extract layout-specific features and textual content to accurately extract financial data points, including revenues, expenses, and net profits. The approach puts several deep learning models in-line, each of which is pre-trained for financial statement analysis. Experimental validation highlights the effectiveness of VRDM in accomplishing end-to-end automation of important information extraction[5] tasks in use cases such as financial analysis and reporting.

Chapter 3

MATERIALS AND METHODS

3.1 LayoutLMv3

Developed to comprehend and handle documents that contain text and layout information, including visual information, LayoutLMv3 [8]is a cutting-edge model. By taking use of the spatial organisation of the text within a document, LayoutLMv3 aims to improve document comprehension. Along with linguistic understanding of documents, it also merges picture and text modalities to boost performance in tasks like reading forms, interpreting receipts, and classifying documents.

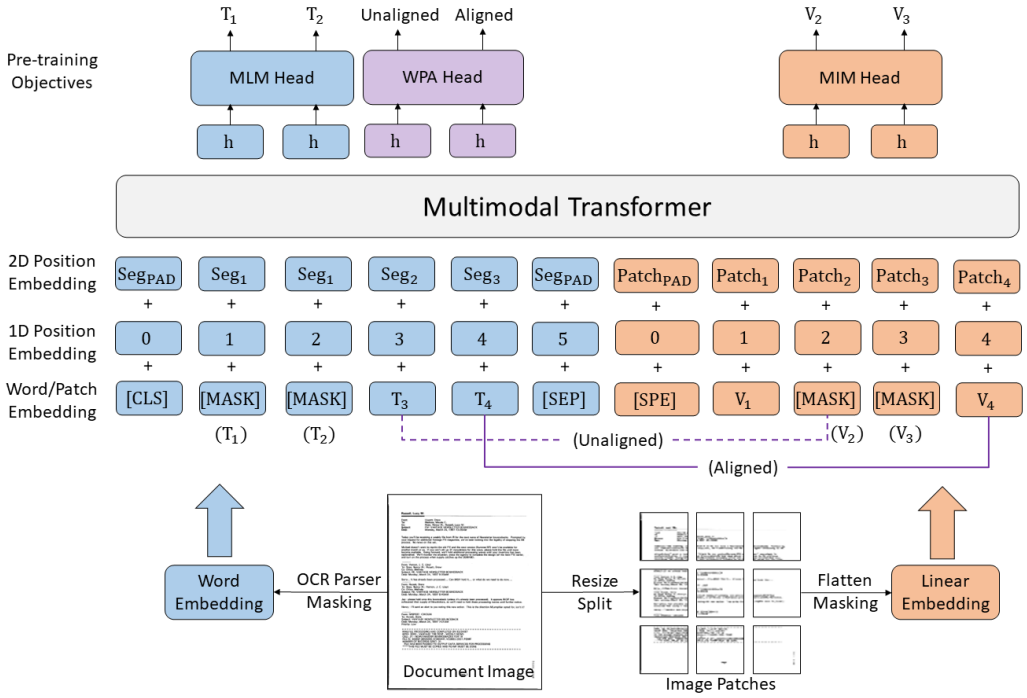

Figure 3.1: LayoutLM-v3 architecture diagram [11]

EMBEDDINGS

Text/Word Embeddings

Text in documents is converted into numerical representations in LayoutLMv3 by text embedding. The model's ability to comprehend and digest text effectively depends on this procedure. For complete document understanding, LayoutLMv3

combines visual information with text embeddings created using sophisticated natural language processing algorithms.

Visual Embedding/ Image Embedding

Processing the document's structure and layout is a part of visual embedding. LayoutLMv3 captures spatial connections and visual elements in the document using visual transformers. These embeddings aid in the model's comprehension of the context that the layout provides, including the locations of text blocks, graphics, and other document components.

Process:

- Linear Embedding: After splitting up images into patches, each patch is given a linear embedding.
- CNN: Convolutional neural networks (CNNs) are used to analyse pictures and produce feature maps that resemble grids.
- Faster R-CNN: Specific regions of interest within the image are identified and embedded.

TRANSORMERS

Multimodal Transformer The multimodal transformer, which incorporates both text and visual embeddings, is the central component of LayoutLMv3. This transformer creates a comprehensive knowledge of the document by analysing the merged data. LayoutLMv3 is able to handle complicated document analysis tasks more correctly because to the multimodal transformer[13] , which takes into account both textual content and layout. This allows the model to exploit both textual and visual information for full document comprehension. [30]

Pre-training Objectives

LayoutLMv3 improves its performance by using many pre-training objectives:

- Masked Word Token Classification: enhances the model's capacity for language understanding by teaching it to anticipate words that are hidden inside the text.

- Masked Patch Token Classification: E improves the model's capacity to anticipate picture regions that are veiled, aiding in the comprehension of the image's structure of any length.
- Origin Image Reconstruction: makes sure the model keeps all of the visual information by reassembling the original picture.
- Masked Region Feature Regression: optimises the model's capacity to recognise and anticipate particular areas inside pictures.

PROCESSING

Document Processing LayoutLMv3 excels in processing a variety of document types by leveraging its integrated text and visual understanding capabilities. It is particularly effective in tasks such as form understanding, receipt processing, and document classification. The model's ability to analyze both content and layout allows it to handle complex documents with varied structures efficiently.

3.2 Faster R-CNN

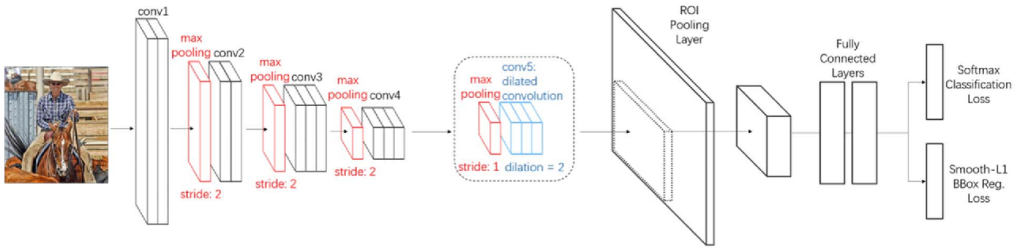

Figure 3.2: Faster R-CNN architecture diagram [10]

LAYERS

Region Proposal Network (RPN) To create a convolutional feature map, the input picture is processed using a deep convolutional neural network (ConvNet), usually a pre-trained model like VGG16 or ResNet. With the ability to capture important high-level visual aspects like edges, textures, and forms, this feature map provides a high-dimensional representation of the input image. These characteristics are essential for the creation of region proposals as well as item categorization.

RoI Pooling Layer

Extraction of fixed-size feature vectors from the suggested regions produced by the RPN is the responsibility of the Region of Interest (RoI) pooling layer. The RoI pooling layer collects features from each mapped region once it has been suggested and mapped into the convolutional feature map. The size of these features is normalised via ROI pooling, which fixes their size so that the ensuing fully linked layers can process them reliably. Because it enables the network to efficiently accommodate region proposals of different sizes, this normalisation is crucial.

Fully Connected Layers

A sequence of fully linked layers are next applied to the pooled features from the ROI pooling layer. The traits that were taken out of the suggested areas are further processed and refined in these levels. The high-level characteristics are consolidated by the fully connected layers, setting them up for the latter phases of bounding box regression and classification.

OUTPUT HEADS

Two parallel branches are involved in the Faster R-CNN final stage:

- Softmax Layer: Over a predetermined set of object classes, this branch generates a probability distribution. The odds that each suggested region belongs to a certain class are output by the softmax layer. This allows the model to correctly categorise the items that are observed.
- Bounding Box Regressor:Modifications to the bounding box coordinates are output by the second branch. By making these changes, the initial area recommendations are improved and the bounding boxes are made to suit the observed items precisely. By adjusting the locations and dimensions of the observed objects, the bounding box regressor increases the localization accuracy.

3.3 Label Studio

An open-source data labelling application called Label Studio offers a flexible platform for annotating many different kinds of data, such as text, photos, audio, and video. When producing the datasets needed to train machine learning models, it is very helpful. Label Studio is essential for annotating documents with key-

value pairs and object identification labels in the context of Faster R-CNN and other document-related tasks.[23]

Key-Value Pair Annotation

Key-value pair annotation is made easier with Label Studio, which is crucial for organising and retrieving particular information from documents. Form processing, and other document comprehension applications frequently employ this kind of annotation. Label Studio aids in the creation of a well-annotated dataset that can be used to train models for tasks like information extraction and form interpretation by labelling sections of text as keys and related values. [25]

Object Detection for Faster R-CNN Labeling

Label Studio offers powerful tools for annotation of pictures with bounding boxes for object detection tasks, which are utilised in the training of Faster R-CNN models. Bounding boxes surrounding interesting things can be drawn by users, who can then identify the boxes. In order for Faster R-CNN to effectively recognise and categorise objects inside pictures, this procedure is essential for producing high-quality training data[24].

Chapter 4

PROPOSED APPROACH

The development of an automated system for document comprehension and sign/seal identification requires a number of intricate procedures.

1. Data Collection and Preprocessing

   Collected a set of documents from the client and conducted an initial review. Filtered out outliers, removed unwanted and duplicate files to ensure a clean dataset.Sorted documents into different categories based on document types.

2. Custom Dataset Creation for LayoutLMv3

   Utilized Label-Studio to create a custom dataset for LayoutLMv3 model training. Annotated documents using Label-Studio's OCR template, capturing key-value pairs for model comprehension.Exported annotated documents in json.min format, totaling approximately 700 files for further processing.

3. Fine-tuning LayoutLMv3 Model

   Fine-tuned LayoutLMv3 using the custom dataset with 10,000 epochs. Studied dataset performance and selected the best model based which is obtained after fine tuning on manual inference verification.

4. Dataset Creation for Faster R-CNN:

   Focused on sign and seal detection, created a dataset by labeling relevant items using Label-Studio.

5. Fine-tuning Faster R-CNN Model

   Fine-tuned Faster R-CNN using the COCO format with 350 epochs and 480 labeled images. Achieved improved model performance suitable for the specific project use case.

The work flow of the project is as shown below:

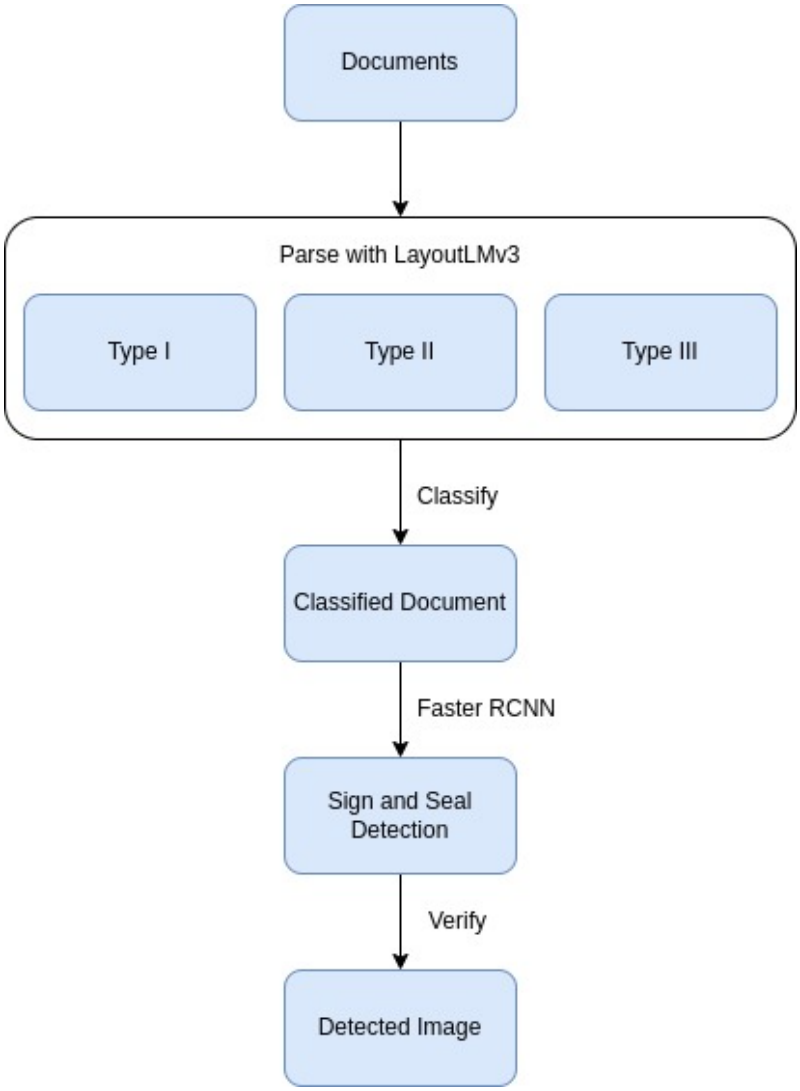

Figure 4.1: Work flow of the project[29]

4.1 COMPARISON OVER OTHER METHODS

1. LayoutLM

   Document understanding problems were revolutionised by LayoutLM, which was first offered with a model architecture that included a BERT-based model with a Masked Region Proposal Network (MRPN). LayoutLM addressed the difficulties presented by documents with complicated structures, such as forms, bills, and receipts, by seamlessly merging text and layout information. Its capacity to effectively extract structured data from documents with a variety

of layouts is its primary invention, which positions it as a fundamental tool in the fields of document processing and OCR (Optical Character Recognition).

2. LayoutLMv2

   With improved design and performance metrics, LayoutLMv2 expands upon the framework created by its predecessor. Precision, recall, and F1-score measures are all higher with the new model architecture than with LayoutLM, allowing it to manage even more complex document layouts with more efficiency. Additionally, LayoutLMv2 includes new training methodologies, namely in pre-training tactics and transfer learning. By addressing the scalability and performance limits seen in previous versions, these enhancements seek to increase the model's flexibility across a variety of document kinds and languages.

3. LayoutLMv3

   At the cutting edge of document understanding technology, LayoutLMv3 has cutting edge improvements in both its architecture and performance. With optimal inference speed, resource utilisation (GPU memory, for example), and overall inference time, this version excels in accuracy and efficiency. Furthermore, LayoutLMv3 may be precisely customised to certain domains or document formats, providing unmatched fine-tuning versatility. LayoutLMv3's enhanced customisation possibilities and smooth integration into pre-existing document processing workflows make it an excellent tool for practical applications.

Thus, I inferred that the transition from LayoutLM to LayoutLMv3 signifies a noteworthy progression in the field of text comprehension technology. Every iteration has improved upon the shortcomings and strengthened the areas of its predecessor. For applications requiring accurate data extraction and strong document layout analysis, LayoutLMv3 becomes the clear choice. It provides improved fine-tuning capabilities, improved model design, and greater performance metrics. I found out that LayoutLMv3 is the best model that can be used within these versions for my project.LayoutLMv3's developments establish it as a fundamental component of contemporary document processing systems, facilitating accurate and efficient document comprehension for a wide range of industries and use cases.

Chapter 5

RESULTS AND DISCUSSIONS

LayoutLMv3 Fine-tuning

The custom dataset generated from the annotated documents was used to refine the LayoutLMv3 model. 10,000 epochs were used in the fine-tuning phase to help the model learn and comprehend the complex linkages and features found in the document layouts. When the optimised LayoutLMv3[27] model's performance was assessed using common metrics, the following outcomes were obtained:

Table 5.1: **Evaluation Metrics and Values for LayoutLMv3**

Metric	**Value**
Eval Loss	0.077
Eval Precision	0.768
Eval Recall	0.763
Eval F1	0.765
Eval Accuracy	0.989

These findings show that the LayoutLMv3 model achieved high precision, recall, and overall accuracy by successfully learning to identify and comprehend the texts' structure and content.

Faster R-CNN Fine-tuning

We employed the Faster R-CNN model for the detection of signs and seals. Between 380 and 400 papers' worth of signs and seals were labelled to provide a different dataset. The COCO format was used to export this dataset, which had JSON files and labelled photos. 480 labelled pictures and 350 epochs were used to refine the Faster R-CNN model. The table below provides an overview of the performance of the optimised Faster R-CNN model:

Table 5.2: **Performance Metrics of Faster R-CNN**

Metric	**SEALS**	**SIGNS**	**Overall**
Precision	0.87	0.82	0.845
Recall	0.83	0.79	0.81
F1 Score	0.85	0.80	0.825
mAP	0.81	0.78	0.795

These measurements show that the Faster R-CNN model is a dependable tool for document verification jobs since it can recognise and categorise signs and seals properly.

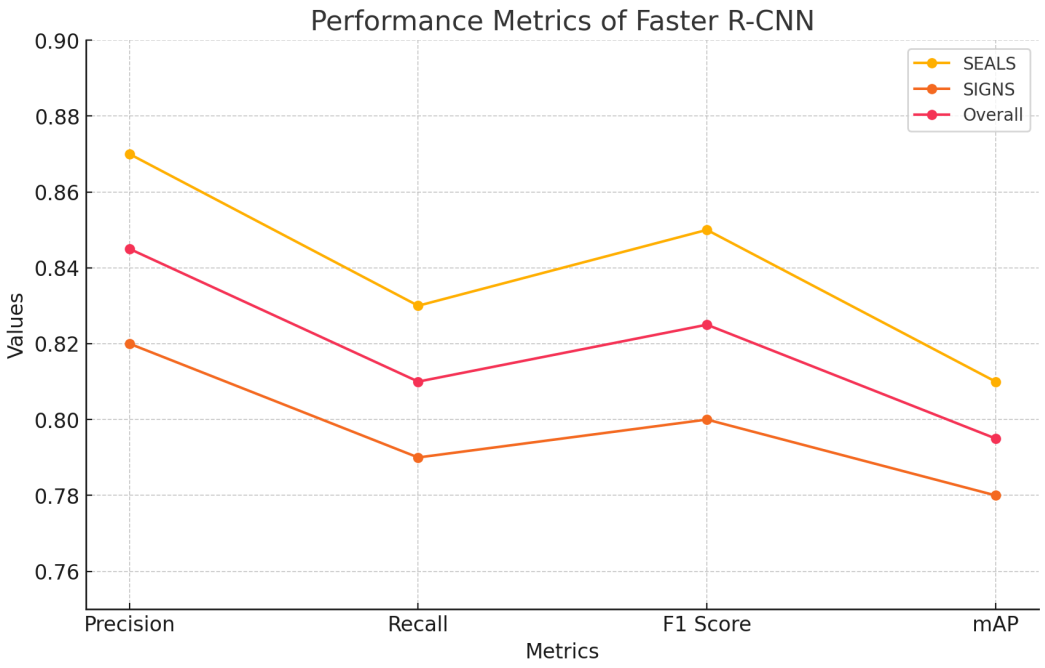

Figure 5.1: Comparison graph of faster R-CNN perfomance matrix

This graph illustrates the performance metrics of a Faster R-CNN model using Precision, Recall, F1 Score, and mAP (mean Average Precision). Each metric is plotted for two categories ('SEALS' and 'SIGNS') as well as an overall average ('Overall'). The lines with markers indicate scores achieved across these metrics, showing how the model performs across different evaluation criteria.

Manual Verification and Validation

After training both models, we manually verified the inferences to ensure their accuracy and reliability. The models demonstrated a high degree of precision in identifying key-value pairs, signs, and seals in the documents. This manual verification step was crucial in confirming the models' capability to perform in real-world scenarios.

Overall System Performance

An automated system for document verification has been created through the combination of LayoutLMv3 for document layout analysis and Faster R-CNN for sign and seal identification. This technology lowers the danger of human mistake, improves operational efficiency, and drastically decreases the labour needed for manual inspection. Now, businesses can rely on this system for reliable, accurate, and consistent document verification—a feature that is especially helpful in settings with large processing volumes. The project's outcomes demonstrate how well-suited sophisticated machine learning models like as LayoutLMv3 and Faster R-CNN are for automating activities related to document verification. Organisations may improve the security, efficiency, and accuracy of their document processing operations by utilising these technologies. These models have been successfully implemented and refined, indicating their potential to revolutionise manual verification procedures and open the door to more dependable and automated solutions across a range of sectors.

Limitations

Despite its name, Faster R-CNN is not the quickest object identification model. Its two-stage detection process can be slower compared to single-stage detectors. The model is complex and can be challenging to adjust, optimize, and implement, consisting of several components, including the Fast R-CNN detector and the Region Proposal Network (RPN), each of which needs fine-tuning for better performance. When using Faster R-CNN for the detection of seals on certificates, some emblems may be mispredicted as seals, indicating difficulty in distinguishing between visually similar elements. While LayoutLMv3 excels at understanding structured documents, it may struggle with documents featuring very different or unconventional layouts, complex formatting, unusual font choices, and a variety of graphical features that differ from the training set.

Figure 5.2: Document type identified by LayoutLMv3 based on the header, and seal and signature detected using Faster RCNN

Figure 5.3: Detected header using LayoutLMv3

Figure 5.4: Detected Seal using Faster R-CNN

Figure 5.5: Detected Sign using Faster R-CNN

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This project successfully demonstrated the development and implementation of an automated document verification system capable of reliably identifying, detecting, and categorizing certificates based on their kind, validating certificates, and rejecting unneeded documents. The system employs a variety of cutting-edge technologies, including LayoutLMv3 for document layout analysis and Faster R-CNN for object detection. The results reveal that the system is capable of extracting data from a wide range of documents with great accuracy and efficiency.

The project's success has far-reaching consequences for businesses that rely significantly on manual document verification processes, which are frequently labor-intensive, time-consuming, and prone to human mistake. By automating this process, enterprises can reduce the burden necessary for manual screening, increase operational efficiency, and improve document security. The system's capacity to effectively detect and classify documents can help to reduce errors and improve document processing quality.

Furthermore, the system's versatility and scalability make it a viable option for a variety of industries, including finance, healthcare, education, and government. The system's flexibility to adapt to various document formats and layouts makes it an invaluable tool for organizations that handle a diverse range of papers. In conclusion, this project has shown how automated document verification systems may change document processing. We can improve the efficiency, accuracy, and security of document verification by incorporating cutting-edge technologies and advanced machine learning algorithms.

6.2 Future Work:

- Dataset Expansion

  Expanding the dataset is essential for improving model performance and generalizability. This can be accomplished by gathering a diversified set of papers from various sectors, domains, and format types. The extended dataset should comprise texts with diverse layouts, fonts, and graphical features, as

well as documents of varying complexity and structure. Furthermore, the dataset should be labeled and annotated with high-quality information to guarantee that the model can effectively learn from it. Data augmentation techniques can be used to expand dataset size while reducing overfitting. Furthermore, active learning and transfer learning can be employed to lower annotation expenses while increasing annotation quality.

- Improved model architectures:

  Improved model architectures can also help to improve performance. Single-stage detectors like YOLO or SSD can be investigated as alternatives to Faster R-CNN, which may be faster and more accurate. Transfer learning can be used to extract features from pre-trained models or to fine-tune them for specific tasks. Hybrid models, which combine the strengths of multiple models, can also be investigated. Ensemble approaches, also known as stacking, can be used to increase accuracy by combining numerous models' predictions. Furthermore, domain adaptation methods can be utilized to modify the model to fit new domains or layouts.

- LayoutLMv3 Enhancements

  LayoutLMv3 is an effective model for comprehending structured texts; yet, it may struggle with uneven layouts or papers with sophisticated formatting. To address this issue, strategies such as layout-aware attention mechanisms can be created to focus on specific sections of the document based on layout data. Self-attention processes can be utilized to represent the links between different parts of the material. Furthermore, domain adaptation techniques can be utilized to adapt the model to new domains or layouts. Layout-aware post-processing algorithms can also be created to improve predictions using layout information.

- Error Correction Mechanisms

  Error correction strategies are crucial for reducing errors and increasing model accuracy. Active learning can be used to choose uncertain samples for human annotation while also incorporating human feedback into the training process. Manual review or re-training are examples of post-processing approaches that can be used to fix faults. Uncertainty estimation approaches can be used to estimate the uncertainty in model predictions and incorporate it into decision-making. Ensemble approaches, such as stacking, can be used

to merge many models to enhance accuracy. By implementing these error correction strategies, the model will become more robust and accurate.

By addressing these areas, we can develop a more robust and accurate automated document verification system capable of effectively identifying, detecting, and classifying certificates with various layouts and structures.[26]

BIBLIOGRAPHY

[1] Wiam Adnan et al. "A LayoutLMv3-Based Model for Enhanced Relation Extraction in Visually-Rich Documents". In: *arXiv preprint arXiv:2404.10848* (2024).

[2] Ammar Nassan Alhajali. *Fine-tune LayoutLM on SROIE Dataset*. Accessed:2024-06-28. 2024. URL: https://www.kaggle.com/code/ammarnassanalhajali/fine-tune-layoutlm-on-sroie-dataset.

[3] Elioenai LG Alves et al. "Information Extraction from Financial Statements based on Visually Rich Document Models". In: *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*. SBC. 2023, pp. 894–908.

[4] Xi Deng et al. "HM-Transformer: Hierarchical Multi-modal Transformer for Long Document Image Understanding". In: *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer. 2023, pp. 232–245.

[5] Mohamed Dhouib, Ghassen Bettaieb, and Aymen Shabou. "Docparser: End-to-end ocr-free information extraction from visually rich documents". In: *International Conference on Document Analysis and Recognition*. Springer. 2023, pp. 155–172.

[6] European Journal of Electrical and Computer Engineering. *Article 596*. Accessed: 2024-06-28. 2024. URL: https://www.ejece.org/index.php/ejece/article/view/596.

[7] Ross Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[8] Zhangxuan Gu et al. "Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4583–4592.

[9] Zhenwei He et al. "Multi-adversarial faster-RCNN with paradigm teacher for unrestricted object detection". In: *International Journal of Computer Vision* 131.3 (2023), pp. 680–700.

[10] https://blog.paperspace.com. *Faster R-CNN architecture diagram*. Accessed: 2024-06-25. 2020. URL: https://blog.paperspace.com/content/images/2020/09/Fig03-1.jpg.

[11] https://huggingface.co. *LayoutLM-v3 architecture diagram*. Accessed: 2024-06-25. 2023. URL: https://huggingface.co/datasets/huggingface/documentation-images/resolve/main/layoutlmv3_architecture.png.

[12] Yupan Huang et al. “Layoutlmv3: Pre-training for document ai with unified text and image masking”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4083–4091.

[13] Lingxing Kong et al. “A Hierarchical Network for Multimodal Document-Level Relation Extraction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 18408–18416.

[14] Farkhad Kuanyshkereyev. “A Flexible and Efficient Approach for Key Information Extraction”. In: (2023).

[15] Anh Duc Le, Dung Van Pham, and Tuan Anh Nguyen. “Deep learning approach for receipt recognition”. In: *Future Data and Security Engineering: 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, November 27–29, 2019, Proceedings 6*. Springer. 2019, pp. 705–712.

[16] Yuchen Li. “Synergizing Optical Character Recognition: A Comparative Analysis and Integration of Tesseract, Keras, Paddle, and Azure OCR”. PhD thesis. University of Sydney, 2024.

[17] Yajing Liu et al. “Unbiased Faster R-CNN for Single-source Domain Generalized Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 28838–28847.

[18] FIMU Docproc Research. *CIVQA-TesseractOCR-LayoutLM Dataset*. Accessed: 2024-06-28. 2024. URL: https://huggingface.co/datasets/fimu-docproc-research/CIVQA-TesseractOCR-LayoutLM.

[19] Google Research. *Extracting Structured Data from Templatic Documents*. Accessed:2024-06-28. 2024. URL: https://research.google/blog/extracting-structured-data-from-templatic-documents/?m=1.

[20] Phil Schmid. *Fine-Tuning LayoutLM*. Accessed: 2024-06-28. 2024. URL: https://www.philschmid.de/fine-tuning-layoutlm.

[21] Evan Segal, Jesse Spencer-Smith, and Douglas C Schmidt. “Evaluation of Generic Deep Learning Building Blocks for Segmentation of Nineteenth Century Documents”. In: (2023).

[22] Shiksha. *Object Detection Using RCNN*. Accessed: 2024-06-28. 2024. URL: https://www.shiksha.com/online-courses/articles/object-detection-using-rcnn/#:~:text=Faster%20R%2DCNN%20combines%20the,the%20robustness%20of%20object%20detection..

[23] Label Studio. *Export Guide*. Accessed: 2024-06-28. 2024. URL: https://labelstud.io/guide/export.

[24] Label Studio. *Image Bounding Box Templates*. Accessed: 2024-06-28. 2024. URL: https://labelstud.io/templates/image_bbox.

[25] Label Studio. *Optical Character Recognition Templates*. Accessed: 2024-06-28. 2024. URL: https://labelstud.io/templates/optical_character_recognition.

[26] Xiaodong Su et al. "Multi-scale object detection algorithm based on faster R-CNN". In: *Business Intelligence and Information Technology: Proceedings of the International Conference on Business Intelligence and Information Technology BIIT 2021*. Springer. 2022, pp. 379–391.

[27] Ryota Tanaka et al. "Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 17. 2024, pp. 19071–19079.

[28] Technogems. *ML-Based 3-Stage Document Verification and Validation System Using AWS*. Accessed: 2024-06-28. 2024. URL: https://medium.com/@technogems/ml-based-3-stage-document-verification-and-validation-system-using-aws-a9f171048e29.

[29] Muhammed Faris Mukthar M V. *Work flow of the Project*. Unpublished Work. 2024.

[30] Yang Xu et al. "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding". In: *arXiv preprint arXiv:2012.14740* (2020).