Structural Information Extraction from Document Images: Addressing Challenges in Layout Analysis, Table Detection, and Classification

Thesis approved by
the Department of Computer Science
RPTU University Kaiserslautern-Landau
for the award of the Doctoral Degree
Doctor of Engineering (Dr.-Ing.)

to

Mohammad Minouei

Date of Defense:	16.01.2026
Dean:	Prof. Dr. Christoph Garth
Reviewer:	Prof. Dr. Didier Stricker
Reviewer:	Prof. Dr. Faisal Shafait

DE-386

Abstract

Paper documents remain a vital part of our daily lives, and the need for automated systems to analyze and extract valuable information from these documents is increasingly important. Recent advancements in artificial intelligence have raised user expectations for the extraction of structural information from document images, going beyond the traditional goal of extracting raw text from documents. Typically, document understanding systems comprise multiple components, including layout analysis, table detection, and document classification, each of which presents unique challenges. These challenges include handling complex and varied layouts, addressing the issue of imbalanced datasets, and developing systems that can adapt and learn over time. Layout analysis is a critical component of document understanding, as it involves organizing and structuring the various elements of a document, such as text, tables, and figures. Accurate table recognition is also essential, as it enables the effective extraction and interpretation of structured data.

This research enhances document analysis by increasing accuracy, robustness, and efficiency, which addresses current shortcomings in structural information extraction from documents through novel datasets, model architectures, and learning strategies. The dissertation presents multiple contributions to the field of document understanding. Initially, we developed a CNN-based method for layout analysis, achieving a 3 percent enhancement over baseline techniques on PubLayNet. Secondly, we introduced a continual learning strategy employing experience-replay techniques, which reduced catastrophic forgetting in table detection by 15 percent. Third, we presented a novel dataset and developed an asymmetric convolution-based neural network, improving table ruling line recognition. To mitigate class imbalance in document classification, we integrated visual and textual features with a customized loss function, resulting in a 13 percent increase in accuracy. The utilization of Large Language Models (LLMs) for document comprehension was also studied. A technique for fine-tuning large language models by structuring input as HTML was created, yielding results on par with state-of-the-art methods while requiring less computational power. And a three-phase prompt engineering strategy for zero-shot information extraction was empirically evaluated, yielding promising outcomes.

Acknowledgement

I am deeply thankful to everyone who supported me during my time as a PhD candidate. I extend my heartfelt gratitude to Dr. Mohammad Reza Soheili for his invaluable guidance and for providing insightful advice at every step of this journey. I am also sincerely grateful to my supervisor, Prof. Dr. Didier Stricker, for believing in my abilities and for giving me the opportunity to work in the great research environment at DFKI.

Contents

1 Introduction

> *One day you will find what you've been looking for, but by then, you won't want it anymore.*
>
> **— Dave Tarnowski**

The digitization of documents has significantly expanded access to information, driving the need for automated methods to efficiently retrieve and analyze data. With advancements in digital imaging, documents now often include both text and images, which together contribute to their meaning. Consequently, modern search engines and information retrieval systems are beginning to integrate methods for extracting information from both modalities, aiming to improve the comprehensiveness and accuracy of search results. Moreover, the growing volume of digital and scanned documents available online has created a pressing need for automated systems capable of efficiently extracting and analyzing information, thereby facilitating knowledge discovery and decision-making.

The analysis and understanding of documents have been a major focus of research for decades. From simple personal tasks like creating shopping lists to complex legal matters, enabling computers to comprehend documents is essential. Unlike humans, who can easily grasp the meaning of a document, computers require a systematic approach. This process, known as document understanding and information extraction from document images, involves multiple steps and techniques. These steps include, but are not limited to, document classification, text recognition, layout analysis, table recognition, and information extraction.

The sequence of processing steps varies across systems, with some prioritizing layout analysis before Optical Character Recognition (OCR) and others employing OCR as an initial step. Both approaches aim to maximize the accuracy and relevance of information extraction. The wide variety of digital and scanned documents—including newspapers, business letters, and technical drawings—demonstrates the growing demand for automation to ensure efficient processing and analysis. Manual extraction is time-consuming and impractical, making automation essential for tasks such as analyzing text references, recognizing layouts, and extracting data.

Document Layout Analysis (DLA) plays a key role by identifying and organizing structural elements such as text blocks, tables, and images, enabling systems to systematically interpret the relationships between these components. This process is crucial because OCR alone, while effective at extracting text, is insufficient for fully understanding the content of a document. DLA provides context by recognizing the arrangement and interactions of elements within a document, facilitating deeper insights into its structure and meaning.

This multi-stage process serves as a flexible framework for adapting document analysis methods to the specific requirements of various tasks and document types. Overcoming key challenges in text recognition, layout analysis, and information extraction is essential to enhancing the accuracy and efficiency of document processing systems.

Advancements in natural language processing (NLP) have significantly contributed to document understanding, particularly in applications that integrate textual and visual data for multi-modal analysis. The development of large language models (LLMs) has further expanded the capabilities of AI in this field, enabling tasks such as text extraction, question answering, and contextual analysis with greater precision and adaptability. These models leverage deep contextual understanding to address increasingly complex challenges in document processing.

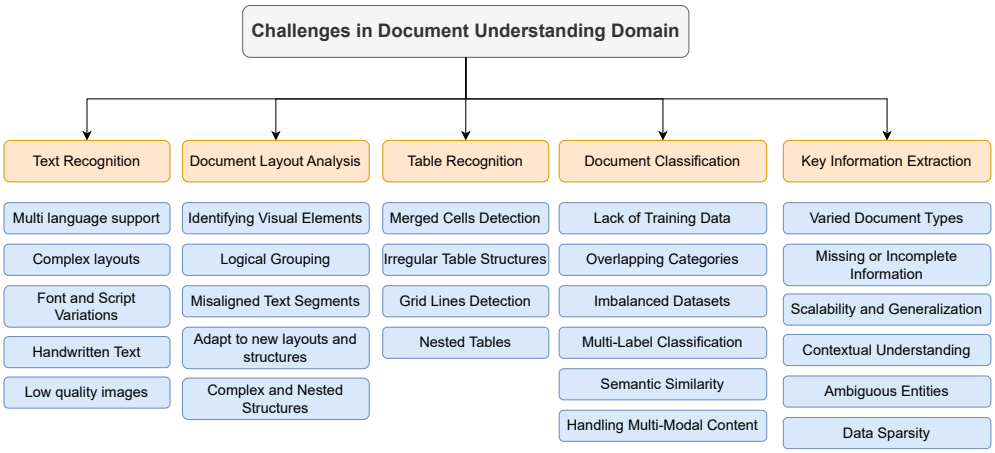


Fig. 1.1: Challenges in Document Understanding: An overview of the key challenges in developing effective document analysis systems, including the need for high-quality, balanced datasets and the ability to handle noisy, incomplete, or imbalanced data.

Figure 1.1 shows the key challenges in the Document Understanding field, highlighting the complexity and scope of problems that must be addressed to develop effective document analysis systems. Document understanding systems face significant challenges because they rely on high-quality, balanced datasets, which are not always

available in real-world settings. This may prevent the training and deployment of effective models because imbalanced datasets can bias classification performance toward majority classes, and limited data availability can limit a model's ability to generalize to new scenarios. To address these issues, advanced techniques for dealing with noisy, incomplete, or imbalanced datasets are required, as well as ensuring the adaptability and reliability of document analysis systems in real-world settings.

1.1 Motivation and Problem Statement

Document digitization has significantly increased access to information, but realizing the full potential of this wealth of data necessitates sophisticated automated systems capable of accurately understanding and extracting meaningful structural information from a wide range of document types, which frequently include text and images. At the heart of this challenge are the inherent complexity and limitations of real-world data, which present significant challenges to achieving robust and accurate document understanding.

Document layout analysis has been studied extensively, with numerous methods proposed over the years. Identifying and segmenting text blocks, titles, figures, tables, and lists in visually complex documents is challenging. Deep learning, especially object detection, has gained popularity in document image analysis. We investigated object detection techniques for document images to improve document layout analysis accuracy and robustness.

Models often encounter new document variations or domain shifts in practice. The need to learn from new data and the risk of catastrophic forgetting must be balanced to maintain model performance while updating it with new data. Catastrophic forgetting affects model stability and requires costly retraining when new information reduces previously acquired knowledge. For robust and reliable document understanding systems, models must be able to learn and adapt to new information without forgetting previous knowledge.

Noise and distortions in scannable images make identifying table ruling lines difficult. High-quality, annotated datasets for granular and specialized tasks like this are scarce, making it difficult to create robust detection algorithms. We used neural networks and asymmetric convolutions to overcome noisy and distorted images and present a new dataset.

While document classification is well-studied, models often exhibit significantly reduced accuracy when trained on imbalanced datasets, a common characteristic of real-world data where certain classes are underrepresented. This imbalance biases models and hinders their generalization capabilities. Our approach improved accuracy by 13 percent on our imbalanced document classification dataset by combining the strengths of both modalities.

The ultimate goal of any intelligent document analysis system is to extract relevant structural information from documents. Although this once seemed impossible, the development of large language models has brought it closer to reality. However, successfully applying these powerful tools to extract structured information from visually complex documents with little or no task-specific training data (few-shot and zero-shot learning) presents unique challenges. Ensuring that these models can accurately interpret layout, maintain formatting, and generalize across unknown document types necessitates novel approaches to input representation and prompting. In our research, we presented a method for fine-tuning LLMs by formatting input as HTML, allowing us to realize their full potential in document understanding.

Adding zero-shot document information extraction, especially for visually rich documents, requires sophisticated Multimodal Large Language Models that seamlessly integrate textual and visual understanding. Developing methods that allow these models to accurately interpret complex layouts and extract structured information without task-specific fine-tuning and only robust prompting strategies is another considerable challenge.

This thesis directly addresses these multifaceted challenges by developing novel deep learning architectures, datasets, and continuous learning techniques to advance intelligent document analysis. From basic layout analysis to sophisticated zero-shot information extraction based on large language models, we aim to improve document understanding systems' robustness, efficiency, and adaptability under imperfect data challenges.

1.2 Thesis Organization and Contributions

This thesis is structured to investigate several key challenges within the document understanding domain. Each subsequent chapter details a specific study, including a background section that explains the relevant state-of-the-art methods.

Chapter 2 Layout Analysis Using Object Detection This chapter proposes a novel approach to document layout analysis based on object detection. This method demonstrates the power of object detection concepts for robust layout analysis. This work's results are published in [1].

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. "Document layout analysis with an enhanced object detector". In: *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE. 2021, pp. 1–5

Chapter 3 Enhancing Table Detection Through Continual Learning This chapter introduces continual learning techniques for improving table detection while preventing catastrophic forgetting. It proposes using a Pyramid Vision Transformer (PVT) as the backbone network and experience-replay techniques to significantly reduce forgetting, achieving a 15 percent reduction in the forgetting effect when compared to traditional fine-tuning methods. This contribution appears in [2].

Mohammad Minouei, Khurram Azeem Hashmi, Mohammad Reza Soheili, Muhammad Zeshan Afzal, and Didier Stricker. "Continual learning for table detection in document images". In: *Applied Sciences 12.18* 2022, p. 8969

Chapter 4 Table Ruling Lines Recognition: This chapter presents a novel method for detecting ruling lines in noisy images, which is supported by the creation of a new dataset, *TabLines*, comprising 35,000 labeled samples. It focuses on segmenting ruling lines in table images and suggests a compact CNN model for the task. The technique's effectiveness is demonstrated on TabLines and published in [3].

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. "Efficient table border segmentation with asymmetric convolutions". In: *Fourteenth International Conference on Machine Vision (ICMV 2021)*. Vol. 12084. SPIE. 2022, pp. 133–140

Chapter 5 Imbalanced Document Classification This chapter describes a multimodal approach to addressing class imbalance in document classification that effectively uses both image-based and text-based features. This work highlights the significant potential of multimodal learning in dealing with real-world data challenges and is published in [4].

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. "Multi-Modal Approach for Imbalanced Document Classification". In: *17th International Conference on Machine Vision (ICMV 2024)*. SPIE. 2024, pp. 133–140

Chapter 6 Structural Information Extraction Using LLMs This chapter proposes a novel approach for embedding layout information within text to enhance document understanding using large language models. This method showcases the robust

capabilities of LLMs in interpreting complex document structures and is published in [5].

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. "Embedding Layout in Text for Document Understanding Using Large Language Models". In: *International Conference on Document Analysis and Recognition*. Springer. 2024, pp. 280–293

Chapter 7 Zero-Shot Document Information Extraction using MLLM This chapter delves into zero-shot document information extraction using Multimodal Large Language Models. It highlights how MLLMs integrate text and image understanding, guided by carefully crafted prompts, to achieve comprehensive document insights that extend beyond unimodal approaches.

Chapter 8 summarizes the key findings of the thesis, reflecting on the significance of the contributions made towards addressing data challenges in document understanding.

2 Layout Analysis Using Object Detection

> *Stop worrying if you'll ever truly be happy.*
> *You won't.*
>
> **— Dave Tarnowski**

2.1 Introduction

An essential part of information retrieval and document processing systems is document layout analysis, which is made to detect and arrange various document components such text blocks, pictures, tables, and figures. Historically, layout analysis relied on rule-based and heuristic approaches. However, the emergence of convolutional neural networks and the advancement of object detection frameworks have enabled more robust and adaptable methodologies [6].

Recent advances in deep learning have significantly improved DLA approaches by allowing document fragments to be handled as objects similar to those found in natural photographs. The effective integration of segmentation and object identification procedures into DLA resulted in more accurate and efficient layout analysis methods, as well as improved document element recognition and localization. DLA has thus become a more reliable instrument for enabling future research such as optical character recognition and semantic interpretation, as well as for assisting in the creation of increasingly powerful document processing and information retrieval systems.

There are essentially three steps to object detection: finding regions of interest, classifying those regions, and last, constructing bounding boxes to pinpoint and delineate the objects. Document layout analysis using this methodology has various benefits, such as efficient processing times, high accuracy in finding and classifying pieces inside complicated layouts, and the ability to recognize a broad variety of document components. Layout analysis has become an indispensable tool for many

applications because to the recent advancements in object detection algorithms that allow for faster and more accurate results.

However, applying object detection to DLA poses challenges, such as intricate layouts with densely clustered elements and the differentiation of visually similar components. To address these challenges, we propose a framework for document layout analysis using object detection techniques. Our methodology has demonstrated 3 percent improvements compared to baseline methods, as evidenced by its performance on the PubLayNet dataset.

This chapter explores the application of object detection in document layout analysis, covering its background, development, and our proposed CNN-based architecture and training approach. We present an enhanced framework for the detection of document elements, which demonstrates substantial improvements in mean average precision on the PubLayNet dataset, surpassing baseline models.

2.2 Background

Prior to the advent of deep learning, document layout analysis relied heavily on traditional image processing techniques. Early systems utilized projection profiles, connected component analysis, and morphological operations to identify regions of interest in scanned documents [7]. These methods typically assumed that documents followed clear structural patterns, such as using whitespace detection to separate text blocks and fixed thresholds to distinguish between images and tables [8].

In addition to these techniques, researchers employed methods like run-length encoding and histogram analysis to detect and separate various layout components. The projection profile technique was particularly effective in identifying horizontal and vertical patterns corresponding to lines of text and columns, respectively. By analyzing the intensity distribution along rows or columns, the system could determine potential borders between components based on significant gaps [9, 10].

Another approach involved using decision trees to identify textual and non-textual regions in an image. For instance, the (X Y) segmentation method employed a decision tree where each page of a document was located at the root, and the leaves represented the final segmented regions. At each node, the tree divided the region into two smaller rectangles, repeating this process until no further divisions were possible [11].

The majority of classical methods relied on the geometric properties of document components. For example, connected component analysis grouped highly correlated pixels, enabling the separation of characters and words. Morphological operations, such as dilation and erosion, enhanced these groups by smoothing noise and completing gaps in the detected areas. Although effective in controlled settings, these approaches struggled with documents featuring non-uniform layouts, overlapping items, or noise introduced by the scanning process [12].

Despite the initial successes of these early approaches, they had significant limitations. The use of predetermined thresholds and hand-coded rules made it difficult for systems to adapt to deviations in document format, language, and style. Non-standard layouts, cluttered backgrounds, or variable font sizes frequently caused segmentation errors. As a result, traditional layout analysis, although foundational for automatic document processing, was largely limited to documents with expected patterns and clean visual characteristics.

One of the significant advances of the period was the shift from purely heuristic approaches to methods that had the ability to learn from data. By learning from labeled datasets, machine learning algorithms could infer complex patterns and relationships in document images. Decision tree algorithms, for example, could automatically determine the most discriminating features for differentiating between the various document elements. Ensemble techniques such as Random Forest also improved classification accuracy by combining the predictions of a set of decision trees, reducing the impact of outlier decisions and noise in the data.

With a learning approach, these systems would also learn new document styles with minimal human effort, a crucial advantage in coping with the increasing variability of digital documents. However, the performance of early machine learning systems still remained bottlenecked by the quality of the features extracted [13]. Hand-crafted features, though useful for certain applications, could not model the complex visual patterns of diverse document layouts. As a result, while machine learning was a significant advancement compared to conventional methods, it also highlighted the importance of creating even more advanced techniques that had the ability to learn feature representations directly from raw image data, ultimately leading to the development of deep learning methods that revolutionized the document layout analysis research area.

2.2.1 Layout Analysis with Deep Neural Networks

The emergence of convolutional neural networks transformed computer vision applications, such as document layout analysis. CNNs can learn hierarchical features directly from raw pixels, removing the need for manual feature extraction. This breakthrough enhanced the robustness and accuracy of layout analysis systems [14].

There have been significant advances in CNN designs in recent years, which have had a direct impact on the performance of document layout analysis systems. Residual networks (ResNets) [15] enable deep network training without vanishing gradients, leading to higher accuracy in complicated tasks like layout analysis.

The use of object detection algorithms into DLA has resulted in exact localization of certain elements within documents. Researchers demonstrated excellent accuracy in detecting and finding document components [6]. Figure 2.1 demonstrates a successful DLA outcome, with each document element appropriately detected and labeled. This highlights the ability of contemporary DLA approaches to manage complex layouts with high precision.

There are two main ways to object detection: one-stage detectors and two-stage detectors. As described in Faster R-CNN [16], two-stage detectors incorporate a two-part process: the first phase creates region proposals, while the second phase classifies and refines these proposals and their bounding boxes. In contrast, one-stage detectors execute object localization and classification simultaneously in a single forward pass of the network. They divide the picture into a grid and predict bounding boxes and class probabilities directly for each grid cell [17]. While one-stage detectors are faster, two-stage detectors are known for their superior accuracy [18]. Successful detectors like Mask R-CNN build upon the Faster R-CNN architecture and extend it with a ResNet backbone and a feature pyramid network (FPN) [19], enabling both object detection and instance segmentation.

Figure 2.2 shows how the Faster R-CNN approach extracts a feature map by passing the input image through multiple convolutional layers. A Region Proposal Network (RPN) is then used to this feature map to select regions with the highest likelihood of containing items and mark them with bounding boxes. These recommended regions are then resized to match their original image size, and the bounding boxes are finally sorted into specific object categories. The RPN, introduced in Faster R-CNN, suggests a collection of prospective areas based on the feature maps. The authors incorporated anchor boxes to accommodate different sizes and aspect ratios. For each pixel in a feature map, a collection of anchor boxes with varying scales and ratios is

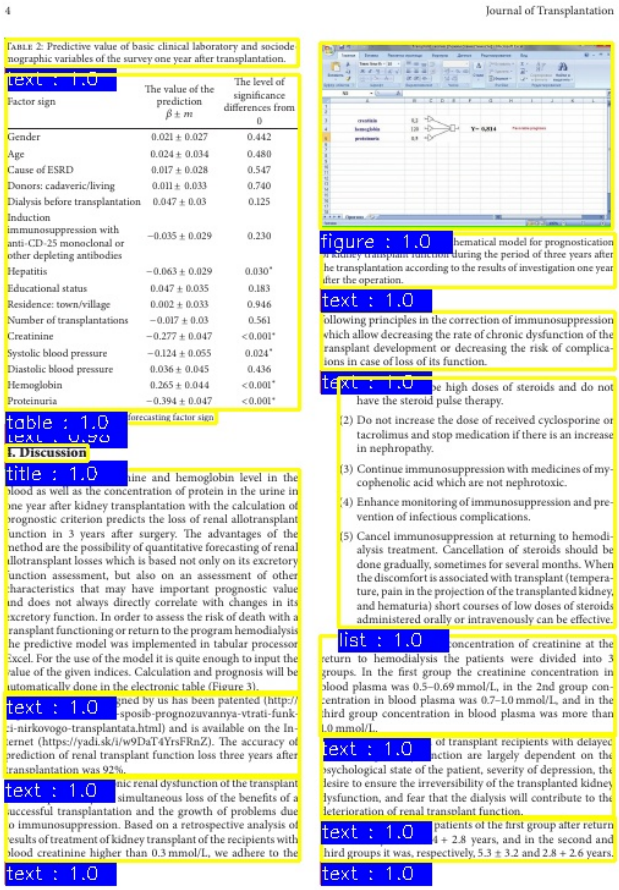


Fig. 2.1: Example of accurate document layout analysis using object detection techniques.

created. These boxes are then classed as foreground or background according to their Intersection over Union (IOU) with the ground truth. This procedure generates hundreds of ideas, and Non-Maximum Suppression (NMS) is employed to eliminate duplicates.

However, the RPN has a drawback in that it uses heuristically specified scale and aspect ratio values for the anchors. This reliance might be a disadvantage. Anchor-free methods use a single anchor to indicate the center of an object [17]. While this methodology is faster, it may not be as accurate as anchor-based methods [18].

Several deep learning models have been presented for document layout analysis that use the object detection paradigm. Yi et al. proposed a CNN-based page object recognition approach with three stages: region proposal extraction, CNN-based classification, and duplication reduction by dynamic programming [20]. Li et al. suggested a hybrid technique that includes linked components, projection profiles,

and conditional random fields for page object recognition, with a CNN applied to refine the classification accuracy of broad regions. Schreiber et al. used the R-CNN object detector to locate tables [21]. They later expanded this architecture to identify tables, equations, and figures [22]. Goswami et al. emphasized the significance of contextual information in document layout analysis by taking into account nearby regions when classifying document elements [23].

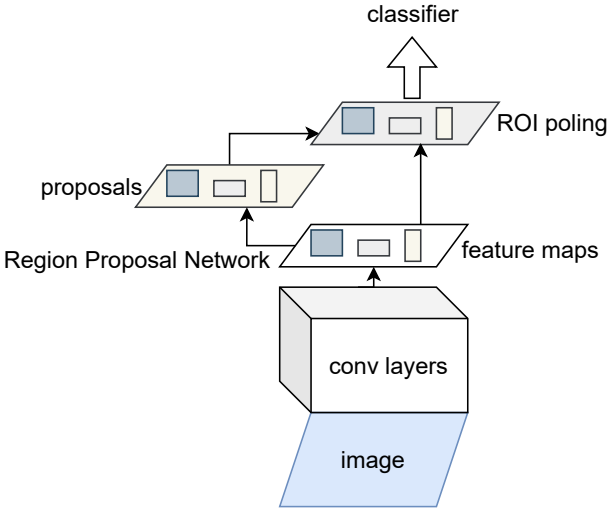


Fig. 2.2: Architecture of a region-based object detection framework. The pipeline begins with feature extraction using convolutional layers, followed by region proposal generation via a Region Proposal Network. The region proposals are refined and processed through RoI pooling to produce fixed-size features, which are then classified and localized by the final classifier.

2.3 Layout analysis with object detection

Objects in images usually change in size, therefore multi-scale analysis is critical for successful recognition and segmentation. Traditional approaches, such as image pyramids, process multiple resolutions of the input image to capture scale changes, but they are computationally expensive and memory-intensive [24]. To solve these limitations, FPN provide an efficient alternative by combining feature maps from different layers of a CNN rather than processing images at numerous resolutions.

Figure 2.3 shows that FPN has two pathways: bottom-up and top-down, with lateral connections. The bottom-up pathway collects feature maps across multiple resolutions from the backbone network, with each stage producing a feature map with

increasingly lower resolution but more semantic information. These feature maps are then routed through the top-down pathway, where higher-level, low-resolution feature maps are upsampled and linked to equivalent lower-level, high-resolution feature maps via lateral connections. FPN creates scale-invariant features by merging high-resolution feature maps, which capture fine-grained spatial information, and low-resolution feature maps, which store semantic context.

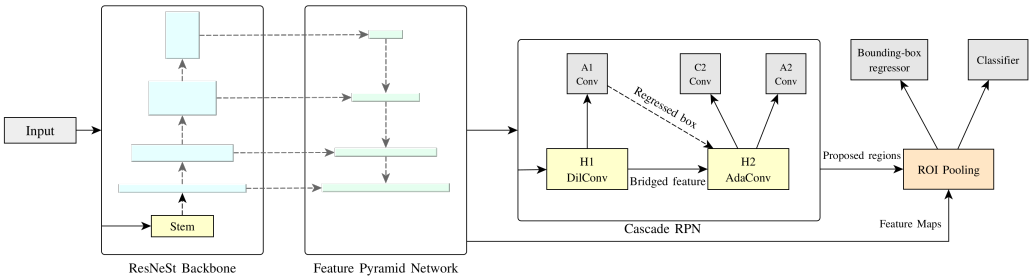


Fig. 2.3: A representation of the proposed method. First, the input image is fed into a CNN. Second, a feature pyramid network fuses the feature maps from the previous step. Third, a cascade RPN is used to find the potential object regions. At last, the ROI-Pooling layer is used to downsize the feature maps and feeds them to both the classifier and bounding box regressor. 'H', 'C', and 'A' designate the head, classifier, and anchor regressor of the cascade RPN.

Feature Extraction

CNNs have proven to be extremely effective at extracting key features from images, securing them as the foundation for tasks such as image categorization and object detection. ResNet [25] is one of the most influential CNN designs. It uses a sequence of convolutional layers and max pooling procedures structured into many stages. Each stage generates feature maps with increasingly finer resolutions, allowing the network to catch both fine-grained and high-level visual patterns.

In our work, we use ResNeSt [26] as the backbone architecture, based on the success of ResNet. ResNeSt improves the traditional residual network by using a multi-path design [27] and a channel-attention mechanism [28] for each path. This approach enables the network to dynamically weight feature channels based on their relevance to the job, resulting in better feature representation. As shown in Figure 2.4, a ResNeSt block is made up of many routes, each with a split-attention mechanism that prioritizes the most informative channels. This capacity allows the network to focus on discriminative features, resulting in more accurate predictions and higher performance across a variety of visual recognition tasks.

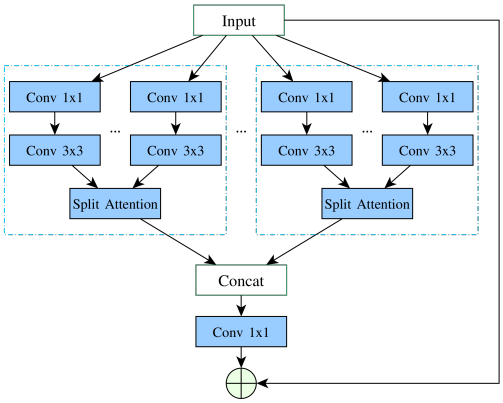


Fig. 2.4: A representation of a ResNeSt block. The input is fed into multiple paths of CNNs. In each path, channel attention mechanism is used. The final output is achieved by concatenating the path's outputs and the input.

Feature Map Fusion

Objects in photos frequently change in size, hence multi-scale analysis is critical for successful recognition and segmentation. Traditional approaches, such as picture pyramids, process multiple resolutions of the input image to capture scale changes, but they are computationally expensive and memory-intensive [24]. To solve these limitations, FPN provide an efficient alternative by combining feature maps from different layers of a CNN rather than processing images at numerous resolutions.

Figure 2.3 shows that FPN has two pathways: bottom-up and top-down, with lateral connections. The bottom-up pathway collects feature maps at various resolutions from the backbone network, with each stage producing a feature map with increasingly lower resolution but more semantic information. These feature maps are then routed through the top-down pathway, where higher-level, low-resolution feature maps are upsampled and linked to equivalent lower-level, high-resolution feature maps via lateral connections. FPN creates scale-invariant features by merging high-resolution feature maps, which capture fine-grained spatial information, and low-resolution feature maps, which store semantic context.

Region proposal network

The Cascade RPN [29] improves on traditional RPN by implementing a multi-stage refinement process. Unlike traditional RPNs, which create region proposals in a single step, Cascade RPN iteratively refines proposals via a series of adaptive

convolutional steps. This approach ensures higher-quality region recommendations by gradually increasing the accuracy of bounding box forecasts. In the first stage, a collection of anchors is evenly initialized throughout the image, with one anchor per place. The second stage refines these preliminary ideas with an adaptive convolution kernel that dynamically modifies the sample area based on the proposed bounding box. This method allows the network to extract more contextual information from the feature maps, resulting in more accurate localization.

The Cascade RPN is especially well-suited for document layout recognition because of its ability to handle non-overlapping rectangular sections effectively. Documents are often made up of structured pieces (such as text blocks, graphics, and tables) that are easily represented by rectangular bounding boxes. Cascade RPN effectively recognizes and refines these elements using a single anchor per place and an adaptive convolution kernel, making it an excellent choice for document analysis jobs.

Object classification

Once the RPN generates a set of region proposals, the next step is to classify these regions into predefined categories. To handle regions of varying sizes, a RoI Pooling layer is employed to resize the feature maps into fixed-dimensional vectors. RoI Pooling is a key component in object detection pipelines, enabling the extraction of fixed-size feature maps from variable-sized region proposals. These vectors are then passed through fully connected layers to predict class labels and refine bounding box coordinates.

The network returns the spatial coordinates and dimensions of the detected bounding boxes, as well as a probability distribution across the available classifications for each region. This end-to-end trainable architecture is designed for document pictures, ensuring high performance in recognizing and classifying elements. The next section describes how the suggested technique was trained, tested, and evaluated using the PubLayNet dataset.

2.3.1 Implementation Details

Our method is developed using the MMDetection codebase [30], which is an open-source object detection toolbox based on PyTorch. MMDetection has a modular design, allowing for extensive modification of detection frameworks, and it offers cutting-edge implementations for a variety of applications, including object detection,

instance segmentation, and panoptic segmentation. The network input is scaled to a maximum dimension of 704 pixels to achieve a balance between computational efficiency and feature resolution. We use the ResNeSt-50 backbone without pre-training and train the complete network on the PubLayNet dataset using eight GPUs. The training configuration includes a batch size of four images per GPU and a mini-batch size of 32, which ensures that hardware resources are used efficiently.

The training procedure utilizes synchronized batch normalization (SyncBN) [31] to stabilize training across multiple GPUs, and stochastic gradient descent (SGD) as the optimizer. The model is trained for 12 epochs, with the learning rate adjusted on the sixth and ninth epochs. The initial learning rate of 0.02 is decreased by a factor of 0.1 at each adjustment point, allowing the model to converge more efficiently. Furthermore, the weight decay and momentum parameters are set to 0.0001 and 0.9, respectively, to regularize the model and improve optimization.

MMDetection's evaluation tools, such as the 'tools/test.py' script, are used to calculate the mean average precision (mAP) and other metrics, including precision, recall, and F1 score. These tools enable COCO-style evaluation, which ensures compatibility with frequently used benchmarks. The combination of MMDetection's modular design and advanced capabilities, such as adjustable learning rate scheduling and SyncBN, allows for robust and scalable training, making it an excellent candidate for document layout analysis.

2.4 Evaluation and Results

2.4.1 Dataset

In the past, datasets for document layout analysis were introduced through ICDAR competitions [32, 33, 34, 35]. However, these datasets often lack the scale needed to train deep CNNs efficiently. The publication of PubLayNet, a large-scale dataset targeted for document layout research, overcomes this problem [36]. PubLayNet, created by Zhong et al. in 2019, has 358,353 annotated document pictures sourced from scientific journals, including both single-column and two-column styles. The dataset includes annotations for five different document element categories: text, title, figure, table, and list, giving it a comprehensive baseline for assessing deep learning approaches in document understanding tasks.

PubLayNet has three subsets: training (335,703 samples), development (11,245), and testing (11,405 samples). The ground-truth annotations for the training and

development sets are made public, however the test set annotations are kept private for benchmarking purposes. Each annotation contains accurate bounding box coordinates and related class names, allowing for supervised training of object identification models. The dataset's variety of document layouts and element types makes it an excellent resource for creating and testing sophisticated document analysis tools.

2.4.2 Evaluation Metrics

As a sub-class of object detection, our task is evaluated using the same performance criteria. The following are the definitions of the common evaluation metrics:

- **Precision** The precision is calculated as the number of true positives (TP) divided by the total number of positive predictions (TP + FP). This metric determines the accuracy of a model. The precision is calculated using Equation (2.1):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

- **Recall** The recall, also known as sensitivity, indicates the rate of missed positive predictions. It is calculated by dividing the correct positive predictions (TP) by all positive predictions (TP + FN). The mathematical definition is shown in Equation (2.2):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

- **Precision-Recall Curve** The precision-recall (PR) curve plots precision versus recall for all possible thresholds. A good object detector should have a high recall rate as well as a high precision rate.

- **Intersection Over Union (IOU)** The IOU measures the overlap between a predicted bounding box and the correct one for an object. It is calculated using Equation (2.3):

$$\text{IoU(A,B)} = \frac{\text{The Overlapping area}}{\text{The Union area}} = \frac{|A \cap B|}{|A \cup B|} \tag{2.3}$$

- **Mean Average Precision (mAP)** The mean average precision (mAP) is a widely used parameter for evaluating object detection models. It is the area under the precision-recall curve for each class, and the mAP is computed by

averaging all average precision values for all classes. Equation (2.4) formulates the metric:

$$\text{mAP} = \frac{1}{N} \sum_{r=1}^{N} AP_r \tag{2.4}$$

where AP_r is the average precision for class r.

For evaluation, we use the IoU metric to measure the overlap between predicted bounding boxes and ground-truth annotations. The performance of a model is quantified using mAP, computed over IoU thresholds ranging from 0.5 to 0.95. This evaluation protocol aligns with the COCO challenge [37] and has been adopted by Zhong et al. in their evaluation of Faster R-CNN and Mask R-CNN on PubLayNet [36]. By following this standardized protocol, our method's performance can be directly compared to baseline approaches, ensuring a fair and consistent assessment.

2.4.3 Results

Table 2.1 compares the performance of our method with comparable results from previous studies. Specifically, Zhong et al. [36] reports on the performance of two baseline methods: Faster R-CNN and Mask R-CNN [38]. Our proposed strategy improves accuracy in four of the five document element categories. The mAP results show that titles are the most difficult category to categorize, most likely because of their visual resemblance to text and the dataset's class imbalance. This conclusion is consistent with previous research findings, emphasizing the inherent difficulty in identifying titles from other text parts.

Figure 2.5 shows the recognition results, with annotations for true positives, false positives, and ground truth. Yellow boxes reflect the network's predictions, green boxes are ground-truth annotations, and red boxes are false positives. The findings show that the network successfully recognizes most text regions, while non-text elements like logos and tables are correctly categorized as non-text areas. However, because titles and plain text have similar visual features, it is difficult to separate them. This misclassification is a typical problem in document analysis since titles are frequently visually indistinguishable from other text sections.

Despite these limitations, the suggested architecture recognizes all five groups with satisfactory accuracy. The results, as shown in Table 2.1 and Figure 2.5, support the usefulness of our approach for document layout analysis. Future study could focus on enhancing title classification by including more contextual clues or utilizing advanced feature fusion algorithms.

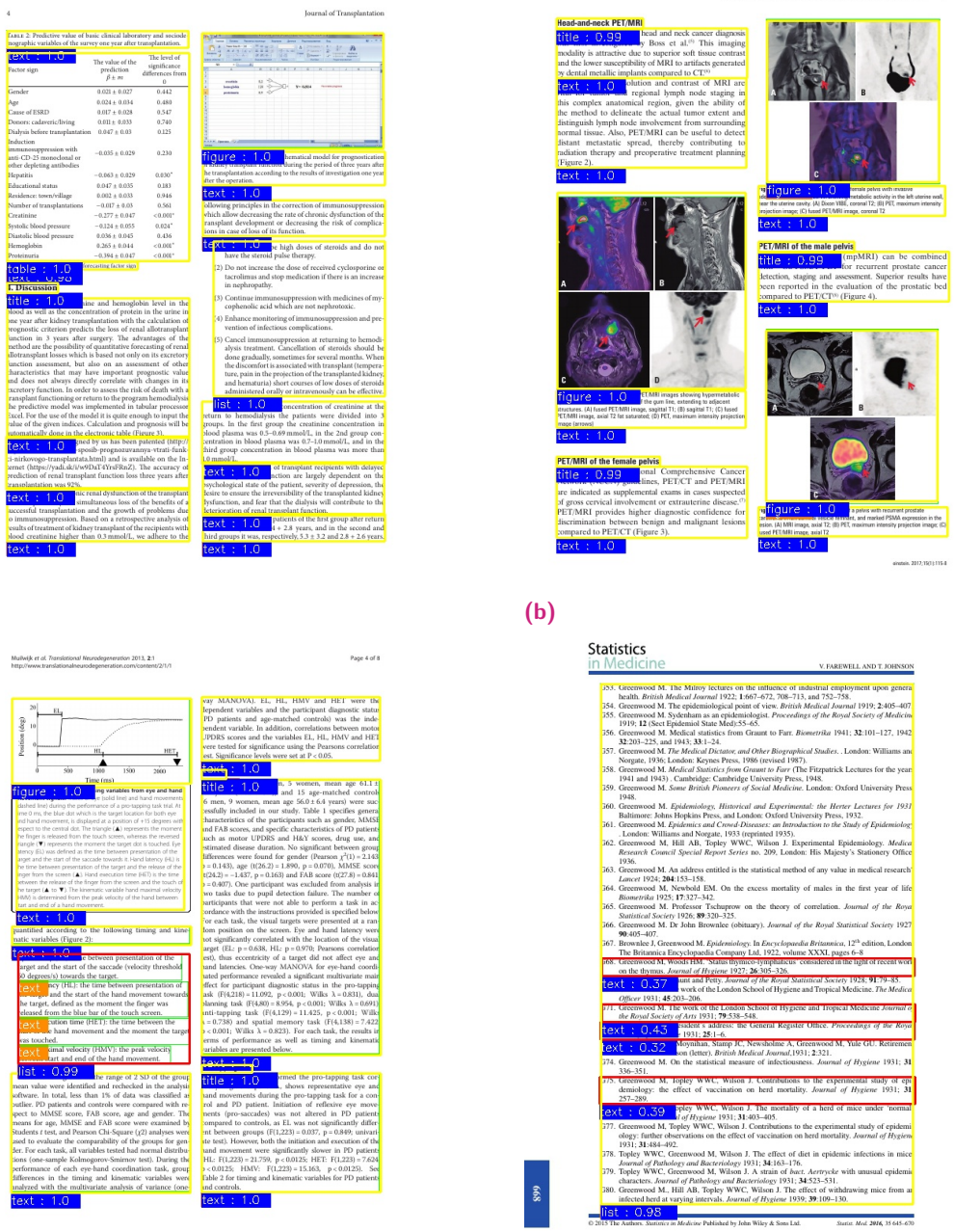


Fig. 2.5: Detection results with corresponding labels. (a) and (b) are true positives, while (c) and (d) have false negatives and false positives.

Tab. 2.1: Results on the validation-set of PubLayNet

Method	AP					Macro average
	Text	Title	List	Table	Figure	
F-RCNN [36]	0.91	0.826	0.883	0.954	0.937	0.902
M-RCNN [36]	0.916	0.84	0.886	0.96	0.949	0.91
CBM [39]	0.886	0.527	0.8683	0.9761	0.8376	0.819
MBC [39]	0.888	0.5279	0.8811	**0.9766**	0.8398	0.8227
Ours	**0.944**	**0.908**	**0.94**	0.974	**0.966**	**0.946**

2.5 Conclusion

In this chapter, we explored various methods for extracting structural information from documents and developed a novel approach for document layout analysis that employs object detection algorithms. The suggested method uses a CNN for feature extraction and a Cascade RPN to identify probable object regions. By combining these components, our system detects document elements including text, titles, figures, tables, and lists in a robust and accurate manner.

To test the proposed approach, we ran numerous tests on the PubLayNet dataset, a large-scale benchmark for document layout analysis. We used the mAP assessment criterion to compare our method to baseline models such as Faster R-CNN and Mask R-CNN. The results show that our methodology enhances the accuracy of document element recognition, exceeding existing methods in four of the five categories. Notably, the adoption of ResNeSt, a robust backbone feature extractor, and the Cascade RPN architecture help to these performance increases by improving feature representation and fine-tuning region proposals.

Our findings have numerous implications for document layout analysis. First, it demonstrates the efficiency of object detection techniques in tackling document understanding issues, especially when combined with advanced feature extractors such as ResNeSt. Second, it emphasizes the importance of Cascade RPN in increasing detection accuracy by iteratively refining region recommendations. These results indicate that our proposed architecture is a promising alternative for activities that need precise and efficient document processing.

Future study could look into additional improvements, such as incorporating more contextual information or using multi-task learning to improve the classification of visually similar categories like titles and text.

3 Enhancing Table Detection Through Continual Learning

> *Good things will always find you, but you'll be too distracted by bad things to notice.*
>
> — **Dave Tarnowski**

3.1 Introduction

The rapid growth of data has created a need for systems that can learn from new data points progressively while preserving previously learned information. However, continuously training a network can lead to a decline in performance on previous data when learning from new instances, a phenomenon known as catastrophic forgetting [40]. This issue is particularly relevant in the context of table detection, a field that has received little attention despite the availability of many datasets and advanced algorithms.

Modern deep learning algorithms have produced excellent results, but they are limited by two major drawbacks: their reliance on large, diverse training datasets and their inability to incorporate new knowledge without damaging earlier learning. The latter restriction is due to neural networks' tendency to overwrite current weight configurations when adjusting to new tasks, as opposed to organic brain systems, which have built-in mechanisms for retaining established knowledge. These limitations are especially problematic in table detection, where the variability of table formats and the need for continuous learning pose significant challenges.

Tables are used in a variety of documents, including scientific journals and financial reports, and their accurate detection and extraction are critical for downstream applications such as information retrieval, data mining, and automated report preparation. However, the inherent variability of table formats—varying boundaries, merged cells, and variable layouts—presents considerable obstacles. To address these challenges, continual learning provides a promising solution, which allows

models to be fine-tuned as new data enters, eliminating the need to retrain from scratch.

Continual learning is especially beneficial in real-world settings, where new document formats and table structures are constantly emerging. By leveraging continual learning, table detection models can adapt to new data and preserve previously learned information, reducing the risk of catastrophic forgetting.

This chapter makes several key contributions:

1. It introduces a novel, end-to-end trainable table detection method that is robust to the inclusion of new datasets while maintaining high performance on previously trained data.
2. It presents the first study to incorporate a continual learning framework for table detection in document images, establishing a baseline for future research in this domain.
3. It highlights the potential of continual learning to address the challenge of dataset-specific training, a major limitation in document image analysis.

3.2 Background

Table detection has been an open problem for several decades. This section discusses earlier rule-based methods, followed by recent deep learning approaches. Finally, we highlight current methods that incorporate continual learning in various fields.

3.2.1 Rule-Based Methods

Early table detection research concentrated on rule-based algorithms that used crafted heuristics based on the physical layout and structural patterns of the tables. Itonori [41] pioneered the use of spatial text-block groupings and ruling lines to identify tables in texts. Chandran and Kasturi [42] highlighted ruling lines as discriminative characteristics, while Pyreddy and Croft [43] created heuristic-driven systems, such as Tintin, to retrieve tables from document images. Green and Krishnamoorthy [44] developed formal layout grammars and geometric models to enhance recognition accuracy.

Coüasnon and Lemaitre [45] and Embley et al. [46] emphasize the limitations of rule-based systems, such as their inability to handle varied table styles and complicated

layouts. Zanibbi et al. [47], Silva et al. [48], and Khusro et al. [49] provide a comprehensive overview of rule-based strategies and their evolution.

3.2.2 Deep Learning Approaches

The introduction of deep learning transformed table detection by enabling data-driven, adaptive solutions. Hao et al. [50] attempted to localize tables by combining CNNs with PDF metadata. Gilani et al. [51] reframed table detection as an object detection problem, utilizing Faster R-CNN [52], a two-stage detector with good accuracy. This method accelerated improvements, notably Siddiqui et al. [53], who created deformable convolutional networks to accommodate geometric variety in tables by adaptively modifying receptive fields.

Instance segmentation improved detection precision in [22], which used Mask R-CNN [38] to generate pixel-level masks for tables. Cascade architectures, such as those by Prasad et al. [54] and Hashmi et al. [55], improved robustness by employing multi-stage detectors with progressively stricter IoU thresholds. In [56], guided table structure recognition through anchor optimization was proposed. Hybrid frameworks, such as Nazir et al.'s [57], combine deformable convolutions with Hybrid Task Cascade [58] to improve performance on irregular layouts.

Alternative approaches investigated fully convolutional networks [59, 60] for pixel-wise classification and graph neural networks [61, 62] to model structural relationships inside tables. Meanwhile, domain-specific adaptations occurred, including Huang et al. [63] employed YOLO architectures for real-time detection, whereas [64] analyzed table structure after detection. TableNet [60] is a hybrid framework that combines structural analysis and detection capabilities. Arif and Shafait [65] improved table detection by combining foreground and background characteristics, employing color coding to distinguish numeric and textual data, and applying Faster R-CNN. Sun et al. [66] proposed a faster R-CNN-based table detecting method that incorporates a corner locating technique.

Table detection has advanced significantly from heuristic algorithms to complex deep learning frameworks. The current state-of-the-art includes deformable CNNs, graph-based models, and hybrid approaches, with ongoing research focusing on improving generalization, handling complex structures, and integrating deeper semantic understanding, as evidenced by comprehensive surveys such as [6, 67].

3.3 Continual Learning

Continual learning (CL) is the process of creating models that can continuously learn from a stream of data, adjusting to new tasks while preserving information from previously acquired tasks. This method differs greatly from standard machine learning, in which models are trained once on a fixed dataset and then kept static. The fundamental issue connected with CL is catastrophic forgetting, which occurs when the model forgets previously learned tasks when trained on new data, as well as the capacity to properly transfer information across tasks. To address these issues, researchers have proposed a variety of methodologies, including regularization techniques, memory-based approaches, and architectural changes such as dynamic network expansion, to enable efficient and flexible learning.

Many researchers have studied ways to learn from new data over time [68]. In the literature, different names are used to describe these techniques, such as life-long learning, incremental learning, and online learning, but here we refer to them as continual learning. The idea of CL is to preserve the knowledge gained from previous training while learning from incoming data [69]. Figure 3.1 illustrates the fundamental difference between conventional approaches and our proposed method that leverages CL. Retraining a model on new data often results in lower performance, but a CL method preserves prior knowledge while learning new information.

Catastrophic forgetting has been tackled by researchers with various techniques in different domains [70, 40]. Some works apply regularization techniques to different parts of the model, such as the loss function, learning rate, and optimizer, while others practice dynamic architecture or parameters isolation for learning different tasks continually. Rehearsal processes have also been exercised [71]. For image classification, Kirkpatrick et al. [72] introduced the elastic weight consolidation (EWC) to alleviate forgetting. In their approach, any modification to the important weights of the network is penalized. In [73], the authors present an EWC-based method for incremental object detection. According to their findings, when the annotations of the old classes are missing, EWC misclassifies previous categories as background. Therefore, they proposed the pseudobounding process for annotating old classes on the new set of images.

Memory based methods are among the most successful methods in this domain. Rebuffi et al. [74] proposed iCaRL for image classification. In iCaRL, a set of exemplars for each class is selected dynamically and used for replay with a knowledge distillation technique [75]. In [76], authors proposed deep model consolidation (DMC) for incremental learning that can be applied in both image classification and

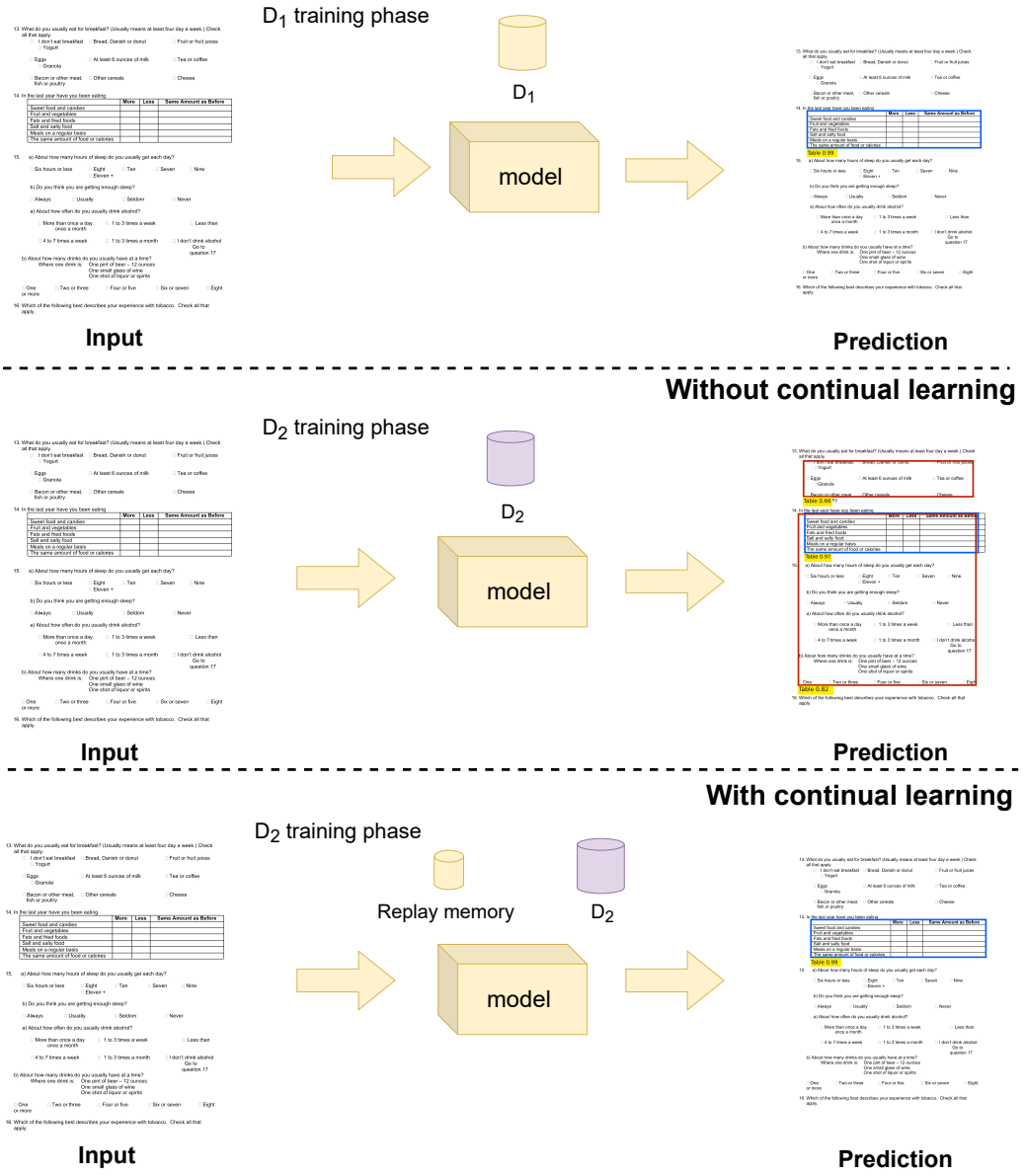
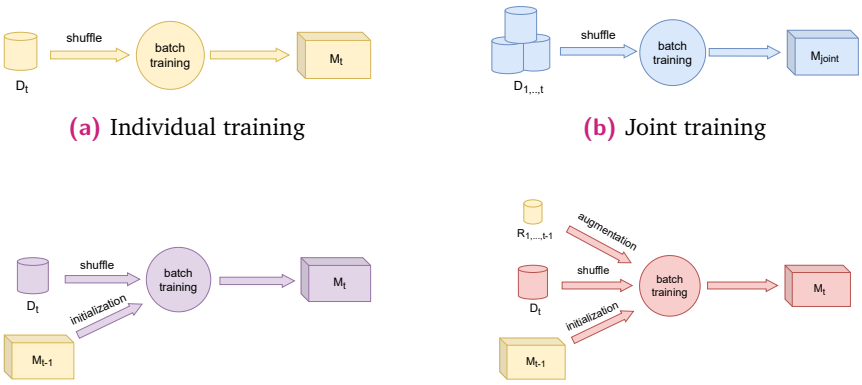


Fig. 3.1: Illustrative sketch of the continual learning usage. After the first training phase (**top** row), conventional methods take the same approach for the next datasets (**middle** row). CL methods can involve previous knowledge in the latest trains by replaying them for the model (**bottom** row). Blue represents true positives, and red denotes false positives.

object detection. In their approach, a double distillation loss is used to combine the two models that one is trained on the old classes and the other is trained on the new ones.

In [77], authors developed a variant of Fast R-CNN [78] with a class agnostic region proposal [79] for object detection in a class incremental setting. A distillation loss



(a) Individual training **(b)** Joint training

(c) Fine-tuning **(d)** Experience replay

Fig. 3.2: The training setup. (**a**) In the individual training approach, the model is trained on a new dataset. (**b**) In the joint training, the model is trained on all available datasets. (**c**) In the fine-tuning approach, the model is trained on a new dataset with the initial parameters obtained from training on the previous datasets. (**d**) In the experience replay approach, at first, the model is initialized with the parameters attained from the previous learning stages on the former datasets; then, the model is trained on a new dataset and a replay memory (that is randomly selected from the former datasets).

was also used to reduce forgetting when training on the new objects with a frozen copy of the learned model on the previous set of classes.

RODEO [80] applied the experience replay procedure using a buffer memory comprising of compressed representations of the past samples. Another recent work that applied experience replay is presented by Shieh et al. [81]; in their method, images from the former task are concatenated with the new samples for class incremental object detection.

3.4 Proposed Approach

The purpose of this study is to continuously train a network with the new data while preserving prior knowledge. In recent years, multiple datasets have been published for table detection with existing demands for more labeled data; thereby, we define the continual learning for table detection as follows. Suppose $D_{1,2,\ldots,t-1}$ is an array of multiple datasets, and M_{t-1} is a model that has been trained on them. At the event of introducing a new dataset at time, t, different scenarios are possible. Figure 3.2 displays four of the possible ways of consuming the new dataset. In the following, we will describe them along with our proposed experiments.

Independent Training

This is the conventional method of training in which a model is trained on each dataset. The Algorithm 1 shows the straightforward batch training procedure which is used here. The results of this experiment will show the upper bound of possible learning with current data and architecture. Figure 3.2a presents this training process.

Joint Training

In the joint training, all the available datasets are exploited. This setup acts as an upper bound of the learning capability of the model using all the available data. As presented in Figure 3.2b, all the available samples are shuffled before the batch training.

Fine-Tuning

The classical fine-tuning procedure is implemented for this experiment. As Figure 3.2c shows, during the training of D_t, a pre-trained model on previous datasets, M_{t-1}, is employed for initializing the parameters of the model. Afterwards, the model is retrained with a lower learning rate on the new instances. Since this setup will result in catastrophic forgetting, its performance is the lower bound for the learners.

Experience Replay

The last experiment is the continual learning technique that we devised for our task, called *experience replay* (Figure 3.2d). In this approach, $R_{1,2,\dots,t-1}$ is a small memory that is dedicated to images of the prior datasets. These images are then presented to the model while the model is trained with the new data. To be precise, each batch contains samples from both D_t and $R_{1,2,\dots,t-1}$. This does not only make the learning more stable but also helps the network strike a balance between learning about new table styles and preserving existing ones.

Algorithm 2 describes our batch training with experience replay. Assume that the training procedure is supposed to proceed for D_t and we have the prior data and the trained model at our disposal. The algorithm begins by initializing the replay memory, $R_{1,2,\dots,t-1}$. It is a random selection of images from $D_1, D_2, \dots, D_{t-1}$. In each iteration of training, a mini-batch is chosen from the D_t and another from

$R_{1,2,\ldots,t-1}$. These batches are then concatenated in one batch, and one step of gradient descent is taken by them.

The number of images in $R_{1,2,\ldots,t-1}$ will be equal to one percent of the number of training samples in D_t. In this manner, we can ensure that the memory is neither too small nor too large for preserving the past knowledge while learning the new ones. Its images are selected randomly from D_is with respect to their size. If s_{D_i} designates the number of training samples in dataset D_i, then the number of images from dataset D_i that reside in $R_{1,2,\ldots,t-1}$ are achieved from (3.1) and represented by C_{D_i}:

$$C_{D_i} = \lceil \frac{s_{D_i}}{\sum_{j=1}^{t-1} s_{D_j}} \times \frac{1}{100} \times s_{D_t} \rceil \tag{3.1}$$

Algorithm 1: Batch training

Input: Learning rate: ν_k, Initial weights from ImageNet

Data: Dataset D_t, batch size bs

Function `BatchTraining(`D_t, bs`):`

- **while** *each iteration* **do**
  - $B \leftarrow$ sample a mini-batch from D_t of size bs;
  - $\mathbf{g} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{w}} L(B)$; `// compute the gradient for the current batch`
  - $\mathbf{w} \leftarrow \mathbf{w} - \nu_k \mathbf{g}$; `// update the current weights`

Algorithm 2: Batch training with Experience Replay

Input: Learning rate: ν_k, formerly trained model

Data: Array of prior datasets $D_{1,2,\ldots,t-1}$, new dataset D_t, batch size bs, memory sample size ms

Function `BatchTrainingWithReplay(`$D_{1,2,\ldots,t-1}, D_t, bs, ms$`):`

- $R \leftarrow$ randomChoice$(D_{1,2,\ldots,t-1})$; `// allocate samples from previous datasets`
- **while** *each iteration* **do**
  - $B_n \overset{bs}{\leftarrow} D_t$; `// sample a mini-batch from` D_t
  - $B_r \overset{ms}{\leftarrow} R$; `// sample a mini-batch from the memory`
  - $\mathbf{g} \leftarrow \frac{1}{bs+ms} \sum_{i}^{bs+ms} \nabla_{\mathbf{w}} L(B_n \cup B_r)$; `// stack two minibatches and compute the gradient`
  - $\mathbf{w} \leftarrow \mathbf{w} - \nu_k \mathbf{g}$; `// update the current weights`

3.4.1 Networks

The choice of network architecture is critical to the effective use of continuous learning approaches in table detection. Faster R-CNN networks are widely utilized in table identification due to their robust feature extraction and exact localization of table sections [6]. Furthermore, innovations in network architecture, such as residual connections and attention modules, make these models resilient in continuous learning scenarios. The inclusion of such architectural advances is critical to maintaining high detection accuracy while responding to new streams of data over time.

In recent years, the introduction of transformer-based models has opened up new possibilities for table identification. The Detection Transformer (DETR) architecture [82] uses self-attention to detect long-range relationships in images. This is especially useful in document understanding, since the spatial context within table components might span huge sections of a page. Hybrid solutions that take advantage of the relative characteristics of both CNNs and transformers are also gaining favor since they allow for both local feature extraction and global contextual sensitivity.

To validate the suggested technique, we chose two cutting-edge architectures: Faster R-CNN [52] and Pyramid Vision Transformer [83, 84], as well as Sparse R-CNN [85]. Faster R-CNN is regarded as a classic baseline in many previous studies, hence it was our initial pick. The Sparse R-CNN+PVT architecture was picked as one of the most recent SOTA detectors.

Faster R-CNN+ResNet

Faster R-CNN is a two-stage detector proposed in 2015 [52]. In the first part of this architecture, a deep CNN is used for extracting feature maps from the input image. We employed ResNet-50 [86] for this purpose. Contrary to Fast R-CNN [78], which uses a selective search algorithm to find the region of objects, the Faster R-CNN utilizes a module called region proposal network to approximate the possible locations of each object in the image. Using RPN, Faster R-CNN can effectively cut the prediction time and improve the accuracy. Moreover, the RPN allows the Faster R-CNN to be end-to-end trainable. After detecting the possible region of interests, the feature map of each ROI is fed to a fully connected network consisting of two branches at the final layer. One branch is a softmax layer to predict the class of objects and the other is a box-regression layer for computing the coordinates. The architecture of Faster R-CNN is seen in Figure 3.3.

Sparse R-CNN+PVT

Pyramid Vision Transformer made the first convolution-free object detector possible [83]. As proposed in [84], the combination of PVT with Sparse R-CNN creates a strong end-to-end method for object detection. In PVT, a progressive shrinking pyramid extracts multi-scale features. Similar to traditional CNNs, as the network grows in depth, the output resolution progressively shrinks. Moreover, an efficient attention layer was designed to further reduce the computation cost.

Given the Figure 3.4, the structure of PVT consists of one main stage repeated four times in order to simulate the pyramid approach, which reduces the size of the feature maps. Inspired by the transformer idea in language translators, the input image must be tokenized. Therefore, the mentioned stage in PVT converts the input to some patches as a dictionary of tokens. So, at the first stage, the input image of size $H \times W \times 3$ is split into $\frac{H \times W}{4^2}$ patches; then the patches are flattened, and a linear projection process is applied to the patches in order to attain embedded equivalents. Afterwards, the embedded patches and their positions are input to another block called transformer encoder, which includes a spatial-reduction attention (SRA) layer. Ultimately, by reshaping the output of the transformer encoder block, the feature map F_1 is obtained. By taking a closer look at Figure 3.4, we can observe that by applying the mentioned processes to the output of the previous stage, the new feature maps F_2, F_3, and F_4 are produced. After that, these feature maps are fed into a sparse R-CNN [85] for object detection. Unlike the conventional RPN, which requires thousands of anchor boxes, sparse R-CNN relies on a small set of learnable proposal boxes. These predictions are then refined in multiple stages.

3.4.2 Implementation Details

In every evaluation, both the Faster-RCNN and the Sparse R-CNN were trained using the publicly available MMDetection toolbox [87]. The training procedure and environment were the same for all experiments. The networks were trained on eight GPUs with four images per GPU. In the ER setting, the batch size is also four per GPU, among which one is from the replay memory.

The ResNet-50 [86] is used as the backbone network for Faster-RCNN, and for the Sparse R-CNN, the PVTv2-B2 [84] is chosen. The backbones of the models were pre-trained on the ImageNet [88]. If not mentioned otherwise, all the default configurations are used from the reference implementations.

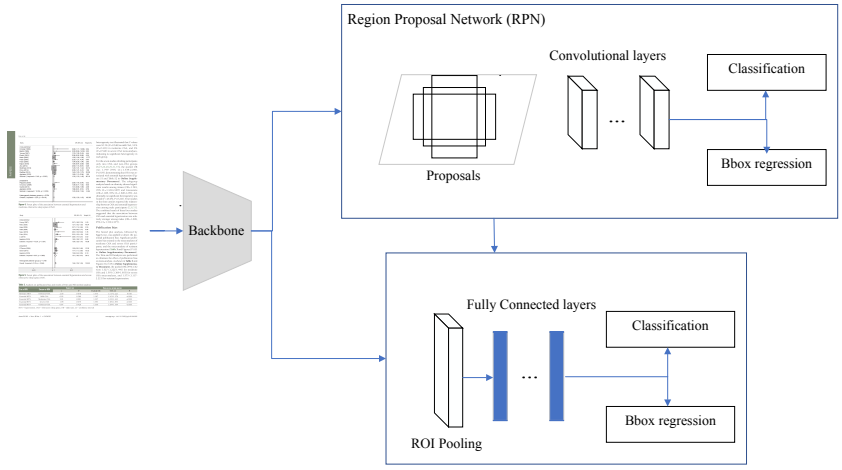


Fig. 3.3: The structure of Faster R-CNN. The workflow consists of feeding the input image to CNN network. Afterwards, potential object regions are found with the RPN. The final step is ROI pooling that extract features for each region and feed them to classifier and the bbox regressor.

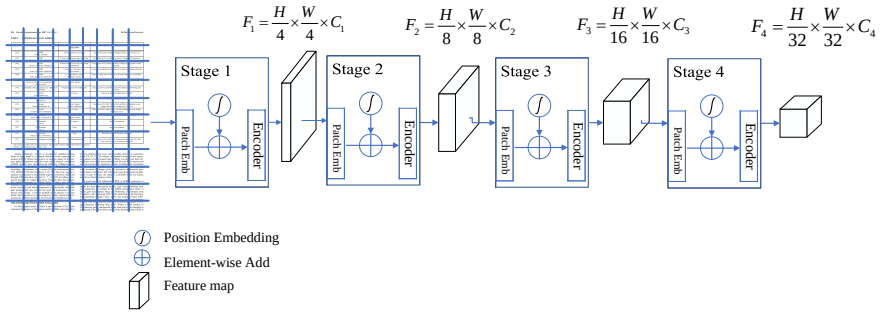


Fig. 3.4: The schema of pyramid vision transformer. Each stage's output is passed to the next layer while the first two dimensions are halved (rows and columns). The finished map will be 16 times smaller than the input yet with a greater depth.

In the IT scenario, the model is trained for three epochs. The initial learning rate is 10^{-4}, which decays with a factor of 10 for every epoch. In FT and ER, the added datasets were fine-tuned for one epoch with a learning rate of 10^{-5}.

To prevent overfitting in the ER scenario, data augmentation is applied on images of replay memory. Four types of augmentation are used from the *Image corruptions* library [89], namely, motion blur, jpeg compression, Gaussian noise, and brightness. The augmentation methods were chosen so that they simulate common real-life scenarios.

3.5 Evaluation and Results

3.5.1 Datasets

In total, we utilized four modern publicly available datasets: TableBank, PubLayNet, PubTables-1M, and FinTabNet. Table 3.1 summarizes datasets' statistical information.

- **TableBank** [90]

  TableBank has been collected from the *arXiv* database [91], containing more than 417 K labeled document images. This dataset comes with two splits, Word and Latex. We combined both for training.

- **PubLayNet** [**zhong2019publaynet**]

  PubLayNet is another large-scale dataset that covers the task of layout analysis in documents. Contrary to manual labeling, this dataset has been collected by automatically annotating the document layout of PDF documents from the PubMed Central™ database. PubLayNet comprises 360 K document samples containing text, title, list, figure, and table. All document samples from the PubLayNet dataset that contain tabular information were excluded for our experiments.

- **PubTables-1M** [92]

  This dataset is currently the largest and most complete dataset that addresses all three fundamental tasks of table analysis. For our experiments, we include the annotations of table detection from this dataset that consists of more than 500 K annotated document pages. Furthermore, we unify the annotations

Tab. 3.1: The number of utilized images in four employed datasets.

Set	TableBank	PubLayNet	PubTables-1M	FinTabNet	Joint
Train	261 K	86 K	461 K	48 K	856 K
Test	8 K	4 K	57 K	6 K	71 K

for various tabular boundaries in this dataset with a single class of tables to conduct joint training.

- **FinTabNet** [93]

  We employ FinTabNet to increase samples' diversity. FinTabNet is derived from the PubTabNet [94] and contains complex tables from financial reports. This dataset comprises 70 K document samples with annotations of tabular boundaries and tabular structures.

3.5.2 Results

In this section, we present the numerical results of the experiments. As outlined in Section 3.4, we conducted four experiments: independent training (IT), joint training (JT), fine-tuning (FT), and experience replay (ER). These experiments were performed using two state-of-the-art models, Faster R-CNN and Sparse R-CNN, for table detection. The evaluation metrics used are the same as those presented in the previous chapter.

In the IT, one model was trained for each individual train-set and tested against the corresponding test-set. In JT, the train-sets of all four datasets are shuffled, and the models are trained on the compiled set of samples. The resultant network is tested separately on the four test-sets of the four datasets. In FT and ER, one network was trained with the four datasets in sequence. Unlike IT, the model will not be tested until it finishes training with all four train-sets. The obtained network will then be tested on the four test-sets with the same order. It is expected that FT approach forgets the first dataset more severely. It should be mentioned that the effect of the order will be studied in a further subsection. As mentioned, the proposed method, ER, takes the same path as FT, with the difference that it consumes a subset of the previous datasets while being fed with the new samples. The four reported results are obtained in a similar manner to FT's. It is expected that ER suffers less from catastrophic forgetting and, ideally, reaches the performance of JT.

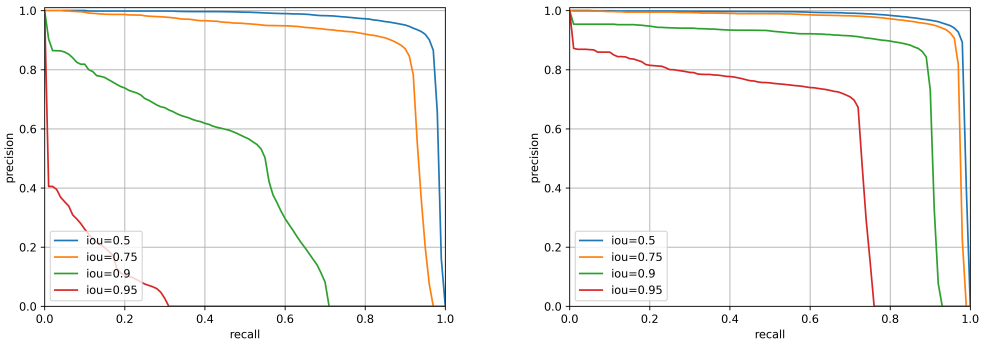
Tab. 3.2: The mAP results of different experiments on multiple test-sets. IT is the Independent training, JT is the Joint training, FT is the Fine-tuning, and ER is the Experience replay. the values written in the parentheses in the ER experiment demonstrate the difference in the mAP metrics between the ER and FT approaches. Acronyms TB, PN, PT, and FN denote TableBank, PubLayNet, PubTables-1M, and FinTabNet, respectively. The R superscripts for ER, demonstrate the index of the previous datasets contributing to the replay memory.

Experiment	Train-Set/Test-Set	Faster R-CNN+ ResNet	Sparse R-CNN+ PVT
IT	$TB/\ TB$	95.7	96.2
JT	$\{TB \cup PN \cup PT \cup FN\}/\ TB$	94.1	94.7
FT	$TB \rightarrow PN \rightarrow PT \rightarrow FN/\ TB$	74.2	76.4
ER	$TB \rightarrow PN^{R_1} \rightarrow PT^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/\ TB$	89.6(+15.4)	90.7(+14.3)
IT	PN/PN	97.6	97.4
JT	$\{TB \cup PN \cup PT \cup FN\}/\ TB/PN$	97.4	97.5
FT	$TB \rightarrow PN \rightarrow PT \rightarrow FN/PN$	90.5	90.6
ER	$TB \rightarrow PN^{R_1} \rightarrow PT^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/PN$	93.7(+3.2)	92.5(+1.9)
IT	PT/PT	98.9	99
JT	$\{TB \cup PN \cup PT \cup FN\}/PT$	98.4	98.7
FT	$TB \rightarrow PN \rightarrow PT \rightarrow FN/PT$	97.2	98
ER	$TB \rightarrow PN^{R_1} \rightarrow PT^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/PT$	97.4(+0.2)	98.2(+0.2)
IT	FN/FN	90	91.3
JT	$\{TB \cup PN \cup PT \cup FN\}/FN$	88.4	92.7
FT	$TB \rightarrow PN \rightarrow PT \rightarrow FN/FN$	90	93.3
ER	$TB \rightarrow PN^{R_1} \rightarrow PT^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/FN$	90	93.1

Table 3.2 summarizes the results of these experiments. By taking a close look at the values of FT and ER, it can be observed that the proposed approach effectively prevents the models from forgetting previous datasets. To emphasize the contrast, parenthesized values in the ER rows show the mAP-gain by ER in comparison to FT. In particular, the mAP for the proposed method on the TableBank is about 15 percent higher than FT. We see that the Sparse R-CNN+PVT demonstrates a better performance than the Faster R-CNN+ResNet in almost all experiments.

The precision–recall curves for ER and FT with the Sparse R-CNN+PVT architecture are presented in Figure 3.6. It is evident that as IOU thresholds rise, FT curves plummet. This is further illustrated with IOU values of 0.9 and 0.95 (in green and red, respectively). Moreover, the fact that the older datasets are more prone to forgetting is again apparent in the figures. This is evident from the TableBank's curves in Figures 3.6a,b and in Figures 3.6g,h that correspond to the most recent dataset.

Figure 3.7 presents common pitfalls ahead of FT. In some cases, the model inaccurately detects the bounding boxes, and in others, we see frequent samples of false-positives. In contrast, the ER approach has led to a better performance and prevented the model from forgetting.



(a) (b)

Fig. 3.5: *Cont.*

3.5.3 The Effect of Datasets Order

It is clear that the inherent differences between the datasets is the cause of catastrophic forgetting. Nonetheless, the results showed that the performance drop of the network is harsher on the older samples. To investigate this, we repeated the experiments in Section 3.4 with a different sequence of datasets during the training phase. In the initial experiments, the order of the datasets was: TableBank, PubLayNet, PubTables-1M, and FinTabNet. However, for the second trial, the sequence is changed to PubTables-1M, PubLayNet, TableBank, and FinTabNet. The rest of the settings are equal to the previous experiments.

The results of this trial are presented in Table 3.3. These results support the previous ones, and the effect of forgetting is once again apparent. By comparing the results of the models on PubTables-1M in the first and second trials (Tables 3.2 and 3.3), we can infer that the performance of models on the preceding datasets drops more drastically. As is presented in Table 3.2, the effect of forgetting on the test-set of PubTables-1M is less than one percent, while in contrast, Table 3.3 shows a 1.1% improvement using ER over FT.

3.5.4 Comparison with State-of-the-Arts

The current SOTA methods heavily rely on particular datasets for training and evaluation. However, in this study, we conducted the experiments on multiple datasets in sequence. To this end, some of the datasets were altered, and the training procedures were different than is customary. Hence, results of this study are not directly comparable to the previous SOTA. Nevertheless, a few of them are reported

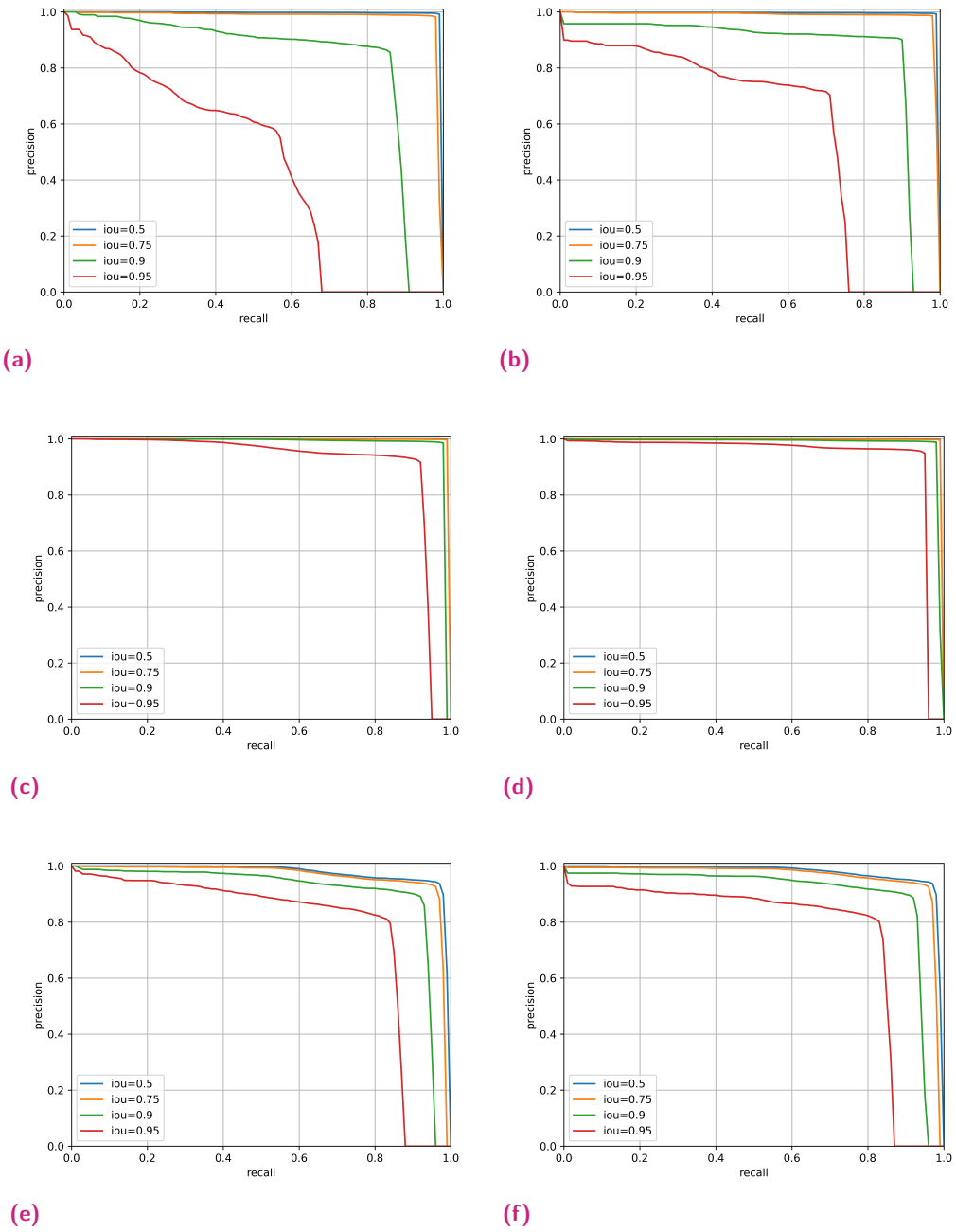


Fig. 3.6: Precision–recall (PR) curves for FT (a, c, e, g) and ER (b, d, f, h) with Sparse R-CNN on four datasets. Each row corresponds to a dataset (from top to bottom): TableBank, PubLayNet, PubTables-1M, FinTabNet. Different IOU threshold are demonstrated with blue, orange, green, and red which correspond to 50%, 75%, 90%, and 95%, respectively.

(a) (b)

(c) (d)

(e) (f)

Fig. 3.7: The qualitative results using two methods: (**a**) Fine-tuning, (**b**) Experience replay. Blue represents true positive, and red denotes false positive. The samples are from TableBank, PubLayNet, and PubTables-1M datasets, respectively. The Experience replay method maintains the performance but the fine-tuning approach suffers from false detection and inaccurate bounding boxes.

Tab. 3.3: The mAP results of different experiments on multiple test-sets. FT is the Fine-tuning, and ER is the Experience replay. Acronyms TB, PN, PT, and FN denote TableBank, PubLayNet, PubTables-1M, and FinTabNet, respectively. The R superscripts for ER, demonstrate the index of the previous datasets contributing to the replay memory.

Experiment	Train-Set/Test-Set	Faster R-CNN+ ResNet	Sparse R-CNN+ PVT
FT	$PT \rightarrow PN \rightarrow TB \rightarrow FN/PT$	96.1	97.4
ER	$PT \rightarrow PN^{R_1} \rightarrow TB^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/PT$	97.7(+1.6)	98.5(+1.1)
FT	$PT \rightarrow PN \rightarrow TB \rightarrow FN/PN$	91.2	93.4
ER	$PT \rightarrow PN^{R_1} \rightarrow TB^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/PN$	94.2(+3)	94.4(+1)
FT	$PT \rightarrow PN \rightarrow TB \rightarrow FN/TB$	76.5	79.8
ER	$PT \rightarrow PN^{R_1} \rightarrow TB^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/TB$	87.9(+11.4)	90.7(+10.9)
FT	$PT \rightarrow PN \rightarrow TB \rightarrow FN/FN$	89.5	92.9
ER	$PT \rightarrow PN^{R_1} \rightarrow TB^{R_{1,2}} \rightarrow FN^{R_{1,2,3}}/FN$	89.1	93

Tab. 3.4: The mAP results of SOTA methods and our continual methods. * indicates that the results are not directly comparable. Acronyms TB, PN, PT, and FN denote TableBank, PubLayNet, PubTables-1M, and FinTabNet, respectively.

Method	TB[mAP]	PN[mAP]	PT[mAP]	FN[mAP]
CDeC-Net [64]	96.5	97.8	-	-
CasTabDetectoRS [55]	95.3	-	-	-
DETR [92]	-	-	96.6	-
Faster R-CNN+ResNet (ours) *	89.6	93.7	97.4	90
Sparse R-CNN+PVT (ours) *	90.7	92.5	98.2	93.1

for the convenience of the reader. Table 3.4 quotes the published values. While our methods trained in a continual setting, their results are close to the conventional methods that had been trained on a single dataset.

3.6 Conclusion

This chapter addressed the critical problem of catastrophic forgetting in deep learning models used for table detection in document images. Catastrophic forgetting, or the loss of a model's performance on previously learned tasks when trained on new data, is a key barrier to developing resilient and adaptive models. To address this issue, the study offered a continuous learning approach based on experience replay, a technique that involves storing a subset of images from previous datasets and replaying them while training new data. This approach efficiently retains knowledge from previous tasks while learning new ones.

The studies were carried out on numerous datasets totaling over 900,000 images, demonstrating the scalability and effectiveness of the suggested method. The study found that the experience replay method reduced the forgetting effect by about 15 percent when compared to standard fine-tuning methods, which are prone to catastrophic forgetting. Notably, the continual learning strategy performed similarly to SOTA approaches trained on single datasets, showing its ability to produce competitive outcomes without requiring costly retraining on complete datasets.

This technique makes it easier to train table detection models with new datasets by reducing catastrophic forgetting. This is a key step toward designing adaptive and resilient table identification techniques that can perform well across multiple domains without requiring retraining from start.

4

Table Ruling Lines Recognition

> *Everything will be okay.**
> **Disclaimer: Results may vary.*
>
> **— Dave Tarnowski**

Automatic table understanding in document images has long been a challenging task in the research community. Tables are a crucial component of documents, containing a significant amount of information. Despite their well-defined structure, which is easily comprehensible to humans, existing OCR methods struggle with table digitization. Although considerable efforts have been made to address this challenge, the problem remains largely unsolved.

Tables occur frequently in printed and digital reports, providing a formatted organization to present data in rows and columns. Table ruling lines detection is necessary to identify the tabular organization, which in its turn is vital for operations like table segmentation, cell extraction, and data mining. Table ruling lines recognition is all about identifying horizontal and vertical lines dividing table cells. This is particularly challenging when table arrangement, paper degradation, scan distortion, and the presence of noise vary.

The process of table understanding is typically divided into multiple stages. The first stage involves detecting tables in an image, known as table detection. The second stage is Table Structure Recognition (TSR), where rows and columns must be identified. Some end-to-end approaches take the input image and output the actual table data, known as table recognition.

One of the primary challenges in TSR is the wide variety of tables, which can differ greatly in structure, size, color, and design. Tables with ruling lines, on the other hand, can make the problem easier to solve. When creating a table, authors frequently draw lines between rows and columns to aid comprehension. These lines provide a clear indicator for structure recognition.

To address this issue, this chapter introduces a specialized Convolutional Neural Network designed for line segmentation using asymmetric convolutions, which can effectively identify and separate ruling lines in tables. Furthermore, a new dataset for table line segmentation is introduced, consisting of 35,000 images with various distortions, which can be used to train and evaluate the performance of the proposed CNN model. These contributions aim to improve the accuracy and robustness of table structure recognition and lay the groundwork for future research in this area.

4.1 Background

Historically, table ruling lines were detected mainly through morphological operations and edge detection techniques. Morphological operations such as dilation and erosion manipulate binary images to improve structural features. Dilation bridges fragmented line segments caused by noise, whereas erosion removes isolated artifacts. Sequential combinations, such as opening (erosion followed by dilation) and closing (dilation followed by erosion), help to refine results. For example, closing can repair torn lines, whereas opening reduces noise. These operations typically use structuring elements (such as rectangular or elliptical masks) whose shapes match the expected line geometries.

Following morphological preprocessing, edge detection operators like Canny and Sobel are used to identify intensity gradients that correspond to table boundaries. The Canny detector applies a multi-stage process involving noise reduction, gradient calculation, non-maximum suppression, and hysteresis thresholding to produce precise edge maps. The Sobel operator, using horizontal and vertical convolution kernels, offers a more efficient but less precise approximation of gradients. These generated edge maps then feed into the Hough Transform [95], which maps image-space points to a parameter space (angle and distance) to detect lines, even fragmented ones. However, this traditional pipeline's effectiveness heavily depends on the initial edge quality; noisy or low-contrast images often result in incomplete or spurious lines. Furthermore, its robustness across diverse document qualities is limited by the sensitivity of various parameters, such as gradient thresholds, accumulator resolution, and minimum line length.

Early research focused on heuristic methods, such as identifying column/row separators via line properties [96] or analyzing text distribution [97]. With advances in deep learning, modern TSR approaches leverage object detection architectures [98, 99, 100, 101, 56] and CNNs for row/column segmentation [102, 103]. For

example, Prasad et al. [101] combined morphological operations with Hough line detection and duplication filtering for bordered tables. Recent surveys [67] highlight these methodologies' growing accuracy but note their frequent neglect of ruling line alignment, leading to suboptimal results when traditional edge detection fails.

Given the complexity of TSR, this work focuses on segmenting the ruling lines in a table image. The outcome of this phase can be used alongside existing techniques to enhance their accuracy. Similar to this work, Lo et al. [104] introduced an efficient model for semantic segmentation that incorporates various techniques such as encoder-decoder, skip connections, and asymmetric convolutions. These techniques will be discussed in more detail in the next section.

4.2 The Proposed Method

This section presents a novel CNN architecture, specifically designed for the task of segmenting ruling lines in table images. The proposed method prioritizes the accurate segmentation of long lines, while minimizing the detection of noisy, small lines that can compromise the overall performance.

A detailed overview of the proposed method is provided below, starting with an introduction to the underlying principles and followed by a presentation of the proposed architecture.

4.2.1 Basic Principles

Encoder-Decoder Architecture

Many successful segmentation architectures, such as SegNet [105] and UNet [106], employ an encoder-decoder style network. The encoder is responsible for summarizing features from the input image into compact feature maps, while the decoder converts these maps back to pixel space to obtain a semantic classification.

Inception Architecture

Inception is a well-known architecture introduced by Szegedy et al. [107]. The design of an inception module features different filter sizes within each layer of convolution. The combination of all the features learned from the different filter

sizes leads to an increase in representational power. Following this idea, we used multiple kernel sizes in our encoder and decoder networks.

Asymmetric Convolutions

Typically, CNNs are built with squared shape convolution kernels, such as a 3x3 kernel. However, a convolution kernel can have any dimensions. Asymmetric convolutions are kernels in which the dimension is not a squared shape, e.g., 1x3. In practice, these types of kernels are used as a low-complexity alternative to square-shaped filters. This is done by combining two consecutive asymmetric convolutions. As shown in Ref.[108], a 3x3 kernel can be replaced by a 1x3 followed by a 3x1 kernel. Later, Peng et al. demonstrated an application of these convolutions in their studies for capturing long-range relations[109]. In Ref. [104], authors employed this approach to create an efficient segmentation model.

4.2.2 Network Architecture

Our network uses asymmetric convolutions to prioritize the segmentation of long lines while suppressing noisy small lines. Figure 4.1 shows the proposed encoder-decoder network architecture. The input image is first fed into a two-layer CNN, the second layer having a stride of two. The first CNN layer's output is then routed through an encoder block, which includes four convolutional paths, two of which use asymmetric convolutions and the others use conventional convolutions. The encoder block's output is then fed into the decoder block, which only uses asymmetric convolutions. To facilitate information flow, skip connections are used to concatenate the blocks' input and output. Finally, a classification layer, implemented as a convolutional layer, is used to generate the final output.

As shown in Figure 4.2, each encoder and decoder block contains multiple types of CNN layers. The encoder block features four convolutional paths, with two using asymmetric convolutions and the other two using traditional convolutions. In contrast, the decoder block only uses asymmetric convolutions. Skip connections, a well-established technique in CNNs, are also employed to concatenate the input and output of the blocks, promoting the flow of information.

The design of our network is motivated by the desire to focus the model on features corresponding to long lines, rather than small, noisy ones. The kernel dimensions for asymmetric convolutions are selected empirically, taking into account the dimensions

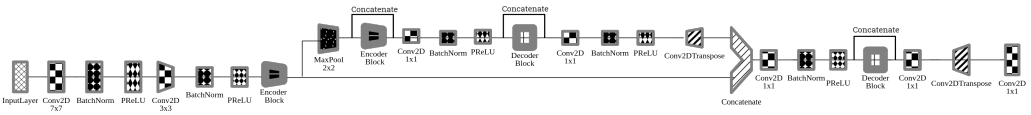
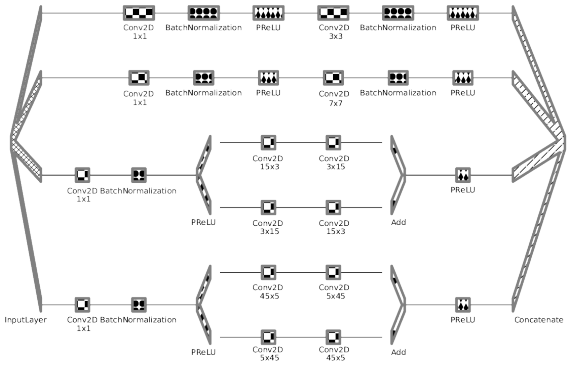
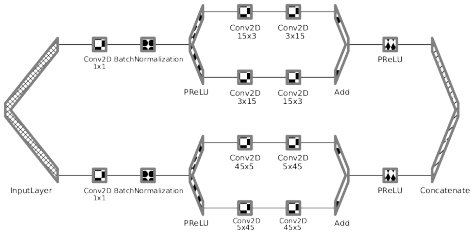


Fig. 4.1: A representation of the proposed method. (All diagrams are made with *Net2vis* [110]).



(a) Encoder block



(b) Decoder block

Fig. 4.2: Encoder and Decoder blocks.

of the feature maps. Specifically, after downsampling twice, the feature maps have dimensions of 32 x 64. The kernels used for asymmetric convolutions are 15 x 3 and 45 x 5, allowing the model to capture both short-range and long-range relationships in the feature map space.

4.2.3 Implementation Details

Training a Convolutional Neural Network with images of varying sizes can be a challenging task. To address this issue, we employed a patch-based approach, extracting patches from the images while preserving their original aspect ratio. To reduce computational time, we resized larger images, ensuring that all images had a dots per inch (DPI) range of 72 to 200.

Given the predominant aspect ratio of greater than one in the dataset, we selected input dimensions of 128x256 pixels for the height and width, respectively. During both training and testing phases, images were divided into patches of these dimensions and fed into the model, allowing for efficient processing and analysis.

The Tensorflow framework [111] is used for the implementation of the network. The batch normalization [112] and PRelu [113] are planted after each convolution.

The model is trained solely on the train set for 4 epochs with the batch size of 8. The SGD optimizer is employed with an initial learning rate of 0.01. The learning rate decays exponentially with a rate of 0.1 after each epoch. Also, the Focal loss [114] is employed to prevent the negative samples from overwhelming the classifier. The training process took 1 hour using a single RTX 3090 GPU.

4.3 Evaluation and Results

4.3.1 Dataset

The DeepFigures dataset [115] is one of the most extensive free-to-use datasets available for graphical page object detection. It comprises over 1.4 million documents, including information on the boundaries of tables and figures. The dataset was created by leveraging scientific articles from public databases such as PubMed and arXiv.

For our experiment, we collected a subset of the DeepFigures dataset and created a new dataset called **TabLines**[1]. To create this dataset, we converted PDF files into images and cropped the table areas. We then altered the text color of the PDF files to white, making only the table lines visible. Finally, we converted the altered PDF files into images and cropped the table areas. The resulting dataset consists of 35,000 images, with 31,000 used for training and 4,000 used for testing.

Data Augmentation

Most datasets for Table Structure Recognition contain only digitally-born images [116, 115, 117, 118]. However, real-world images captured by scanners or smartphones often contain noise and deformation. To simulate real-world scenarios, we applied a series of augmentations to the dataset images using the Imgaug library [119]. This library provides a comprehensive set of augmentations designed for image understanding tasks.

The types and frequencies of the augmentations used are summarized in Table 4.1. All of these distortions were applied to every image in the dataset with a random factor of severity, ensuring that the images are realistic and the trained model is better equipped for real-life scenarios. For geometrical artifacts, both the images and

[1]Available at: https://github.com/minouei-kl/TabLine

the ground truth underwent distortion. For more information on the augmentations used, please refer to [119].

Tab. 4.1: Distortions frequency

Distortion	Probability	Magnitude
Gaussian Blur	0.3	Sigma=(0, 1.5)
Motion Blur	0.3	K=(3, 7)
Average Blur	0.3	K=(1, 3)
Contrast scaling	1	alpha=(0.45, 1.25)
Brightness scaling	0.3	Range=(-30, 30)
Shadow distortion	0.3	Scale=(0.7, 1.3)
Gaussian noise	1	Scale=(0.0, 12.75)
SaltAndPepper	1	P=0.001
Color channel multiply	0.7	Scale=(0.7, 1.3)
Hue and saturation scaling	1	Scale=(0.6, 1.4)
Perspective Transform	1	Scale=(0.001, 0.01)
piece-wise affine	0.7	Scale=(0.0021, 0.0042)

Tab. 4.2 shows the effect of using asymmetric convolution on the model's performance. The total parameters of the model are less than a Million which is fairly lightweight. As presented, replacing the squared shaped kernels with the asymmetric ones results in fewer Parameters and FLOPs. The inference time is computed by averaging the time needed for inferring a single image.

Tab. 4.2: Network parameters, FLOPs (FLoating-point OPerationS), and inference speed.

Method	Params (M)	FLOPs (M)	inference (ms)
without asymmetric	0.88	512	51
with asymmetric	0.86	321	49

4.3.2 Results

In this section, we compare our method with the approach described in Ref. [101]. Since their method is not directly compatible with the segmentation task, we employed a fair comparison strategy. First, we executed their method on the ground truth of the test set and saved the recognized lines as images, which can be considered as the upper bound for their method. We then applied the same steps to obtain

results on the test set images and evaluated these samples against the ground truths using the same metrics as our method.

Our results show that while their method performs well on images without noise, it struggles in the presence of distortions. In contrast, our approach demonstrates improved robustness to distorted images.

To demonstrate the effectiveness of asymmetric convolutions, we conducted an additional experiment where we removed the asymmetric convolutions from the encoder block and replaced them with regular convolutions in the decoder block. Figure 4.3 illustrates a case where the impact of asymmetric convolutions is clear. As shown, the asymmetric method is more successful in preserving long lines and eliminating small noises.

Table 4.3 summarizes the common metrics evaluated on the test set for these three experiments. Furthermore, Figure 4.4 presents samples of the distorted inputs and the outputs of these three methods. It is evident that our network provides plausible results in terms of both qualitative and quantitative measures.

Notably, our findings contrast with those reported in Ref. [104], where the authors found that using asymmetric convolution did not improve accuracy. However, our application differs in nature from their studies, and we are specifically focusing on line segmentation. In our experiments, using asymmetric convolution actually increased the accuracy, highlighting the importance of carefully evaluating the effectiveness of different techniques in the context of specific tasks.

Tab. 4.3: Results

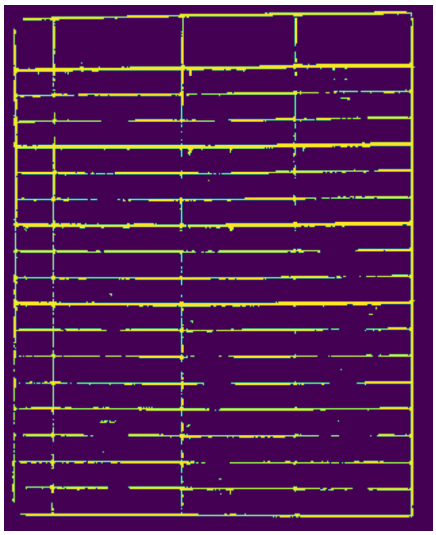
Method	Precision	Recall	F1	Mean IOU
CascadeTabNet[101]	62.450	58.451	60.148	48.896
Ours (without asymmetric)	80.666	84.820	82.691	84.395
Ours (with asymmetric)	**81.163**	**85.787**	**83.411**	**84.954**

4.4 Conclusion

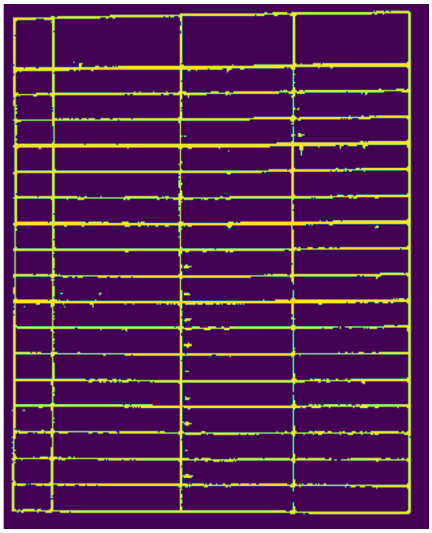
This chapter addressed the challenge of accurately identifying ruling lines in noisy table images, which is a critical step in table structure recognition. Unlike previous methods that rely on hand-crafted features or traditional image processing techniques, we used a proposed dataset, TabLines, which contains 35,000 labeled images

(a) N	(b) Eq. (85)	(c) test results	|(b) − (c)|
4	26.1696	19.815	6.35461
5	17.06	13.42	3.63998
6	10.9365	8.854	2.08247
7	7.24505	5.854	1.39105
8	5.19916	4.265	0.934156
9	4.11975	3.411	0.708747
10	3.56505	2.914	0.651048
11	3.28383	2.727	0.556834
12	3.14225	2.645	0.497247
13	3.07121	2.556	0.515206
14	3.03562	2.543	0.492624
15	3.01782	2.513	0.504817
16	3.00891	2.532	0.47691
17	3.00446	2.508	0.496455
18	3.00223	2.495	0.507228
19	3.00111	2.525	0.476114
20	3.00056	2.496	0.504557

(a) Test image

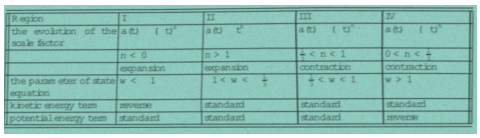


(b) Without asymmetric convolutions



(c) With asymmetric convolutions

Fig. 4.3: The effect of using Asymmetric convolutions.

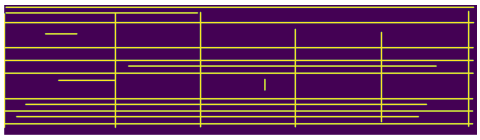


Region	I	II	III	IV
the evolution of the scale factor	$a(t) \quad (t)^{n}$	$a(t) \quad t^{n}$	$a(t) \quad (t)^{n}$	$a(t) \quad (t)^{n}$
	$n < 0$	$n > 1$	$\frac{1}{3} < n < 1$	$0 < n < \frac{1}{3}$
	expansion	expansion	contraction	contraction
the parameter of state equation	$w < -1$	$-1 < w < -\frac{1}{3}$	$-\frac{1}{3} < w < 1$	$w > 1$
kinetic energy term	reverse	standard	standard	standard
potential energy term	standard	standard	standard	reverse

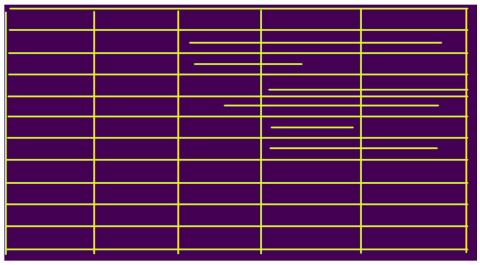
(a)

simplex type	$B^0 \quad F^0$ type	bnd F^0	bnd F	dual graph
[3;1;1]	[4;1]	tetrahedron	triangle rrr	3-val. vertex r
[1;3;1]	[4;1]	tetrahedron	triangle ggg	3-val. vertex g
[1;1;3]	[2;3]	prism	triangle bbb	3-val. vertex b
[2;2;1]	[4;1]	tetrahedron	rectangle grgr	4-val. vertex gr
[2;1;2]	[3;2]	prism	rectangle rbrb	4-val. vertex rb
[1;2;2]	[3;2]	prism	rectangle gbgb	4-val. vertex gb
[2;1;1]r	[3;1]	triangle r	edge r	edge r
[1;2;1]g	[3;1]	triangle g	edge g	edge g
[1;1;2]b	[2;2]	rectangle b	edge b	edge b
[1;1;1]	[2;1]	edge	vertex	polygon

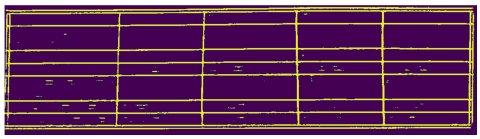
(b)



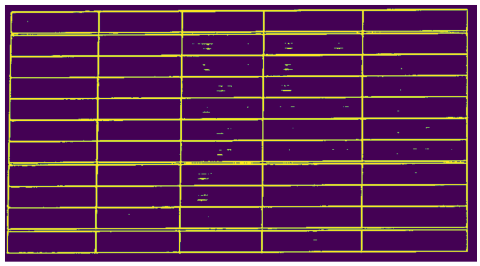
(c)



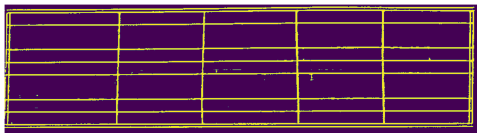
(d)



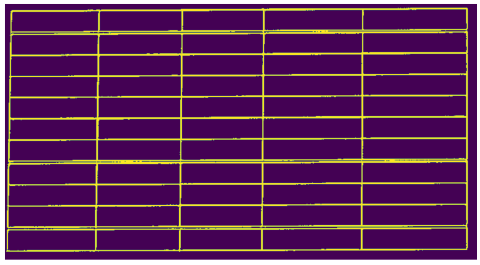
(e)



(f)



(g)



(h)

Fig. 4.4: (a) and (b) are images from the test set. (c) and (d) are the outputs of the CascadeTabNet's line detection method. (e) and (f) are the outputs of our method without asymmetric convolution. At last, (g) and (h) are the outputs of our method.

with various distortions. This dataset addresses the scarcity of publicly available data for training and evaluating ruling line detection algorithms.

We created a custom CNN that uses asymmetric convolutions to improve line detection. Our proposed method outperformed a traditional approach, detecting ruling lines with significantly greater accuracy and robustness, even in the presence of noise and distortions. The network was designed to be lightweight, so it could run on smartphones or be embedded in other networks for real-time line segmentation.

The outcome of this research has the potential to improve the accuracy of table understanding methods that rely on line information, and our approach could be applied to various applications, such as the analysis of ancient documents, where line detection is a crucial step. This approach is more robust than traditional methods and can be utilized in real-time applications, making it a valuable resource for future research in this area.

5 Imbalanced Document Classification

> *You are exactly where you are supposed to be.*
> *Because you make terrible decisions.*
>
> **— Dave Tarnowski**

5.1 Introduction

The issue of data scarcity in deep learning remains a significant, unresolved problem. Many existing works in this domain operate under the assumption that models have access to a comprehensive and balanced dataset that covers all conceivable class conditions. However, real-world scenarios often involve imbalanced and incomplete data, creating considerable challenges. Addressing these challenges is crucial for developing robust models capable of handling diverse and unevenly distributed data.

Documents come in various forms, including administrative paperwork, contractual agreements, research articles, and more. Document classification plays a vital role in efficiently organizing and managing extensive archives of documents for large corporations and organizations, enabling them to streamline their document management processes and ensure effective retrieval. Inspired by the success of Convolutional Neural Networks in image classification tasks, many researchers have applied CNNs to document classification as well. However, document images differ from natural images as they encompass two distinct modalities: textual information and visual information.

Classifying document images based solely on visual or textual information can be suboptimal due to their inherent similarity. To overcome this limitation, researchers have proposed methods that leverage both modalities, integrating textual and visual data sources. This integration has shown significant improvements in the

performance of document classification tasks. By combining textual and visual features, these approaches capture a more comprehensive representation of documents, leading to enhanced classification accuracy and robustness.

In this chapter, we propose a multi-modal approach for tackling the challenge of imbalanced document image classification, which combines both image-based and text-based information to effectively address class imbalance. As shown in Figure 5.1, the approach consists of three main steps. For the visual stream, we employ a two-phase classification strategy using ResNet. During the first phase, the model is trained with cross-entropy loss; in the subsequent phase, the Influence-Balanced (IB) loss function is employed. Next, an OCR engine is utilized to extract textual content from document images, and BERT is employed for text classification. Lastly, the outputs of the ResNet and BERT models are combined, and appropriate techniques are applied to mitigate class imbalance. Our method aims to improve both the accuracy and robustness of imbalanced document image classification tasks. The following sections detail each step, describing the methods used and the rationale behind our approach.

5.2 Background

Imbalanced learning refers to the challenges and methodologies associated with training machine learning models on datasets where certain classes are significantly underrepresented compared to others. This imbalance can lead to biased models that perform poorly on the minority classes.

Existing works in the field of imbalanced learning can be categorized into four groups: balanced data, balanced feature representation, balanced loss, and balanced prediction [120]. These approaches aim to address the challenges posed by imbalanced datasets by modifying the data distribution, improving feature representation, adjusting the loss function, and balancing the final predictions.

Over/under-sampling is a common technique to address imbalanced datasets [121]. However, it can result in over-fitting for tail classes and reduced generalization. It is crucial to carefully consider these drawbacks when making decisions on how to effectively handle class imbalances.

Feature selection methods are employed to improve the model's ability to differentiate between classes by selecting the most significant features. In [122], a set of parametric class-wise learnable centers is proposed. Furthermore, domain-specific

attributes or semantic embedding are incorporated to enrich the feature space in works such as [123, 124].

To tackle imbalanced data, modifying the loss function during network optimization is considered one of the most effective approaches. This allows the model to prioritize minority classes, reducing the bias towards the majority class. Various techniques can be employed to achieve this, such as using weighted loss functions or Focal loss [114].

Another course is to focus on enhancing the inference process. Wang et al. introduced a technique aiming to balance the accuracy of both head and tail classes [125]. They achieve this by training multiple experts concurrently and then combining their predictions through an ensemble approach.

Influence balanced loss

A novel approach for addressing imbalanced datasets presented in [126]. This method, called Influence-Balanced loss, applies re-weighting to samples based on their influence on the decision boundary. The fundamental concept behind the IB loss is to identify the most influential samples that contribute to over-fitting the model in favor of the majority classes and mitigate their impact. This is achieved by estimating the relative influence of a training sample, (x, y), through the calculation of the gradient magnitude with respect to the model parameters W, denoted as $\|\nabla_w L(y, f(x, w))\|_1$. This gradient provides valuable insights into how the loss function changes concerning the model's parameters, allowing for the identification of influential samples during the training process.

Due to the significant impact of changes in the last fully connected layer of the model on the decision boundaries, a weighting factor for IB loss can be formulated as follows:

$$\text{influence weight}(x; w) = \|f(x, w) - y\|_1 \cdot \|h\|_1 \tag{5.1}$$

This equation measures the relative influence of a training sample (x, y) on the model's behavior. It consists of two components: $|f(x, w) - y|_1$ represents the L_1 norm of the prediction error, capturing how much the model's predictions differ from the ground truth labels, while $|h|_1$ denotes the L_1 norm of the hidden feature vector h, signifying the magnitude of the learned features for the sample. By multiplying

these components, the IB loss identifies influential samples impacting the model's decision boundaries and behavior.

The IB weighting factor is then incorporated into the loss function, resulting in the IB Loss (L_{IB}):

$$L\prime(y, f(x, w)) = \frac{L(y, f(x, w))}{\text{influence weight}(x; w)} \tag{5.2}$$

In this equation, $L(y, f(x, w))$ may denote either the cross-entropy or the Focal loss function, which measures the disparity between the predicted outputs $f(x, w)$ and the ground truth labels y. By dividing the loss $L(y, f(x, w))$ by the IB weighting factor, *influence weight*$(x; w)$, the loss is effectively down-weighted for influential samples, especially those coming from dominant classes. This approach helps in addressing class imbalance challenges and leads to improved handling of imbalanced datasets.

To further account for class imbalance, the preliminary practice of down-weighting a sample against its class population can be employed. The weight of a sample will be in inverse proportion with the population of the class that it belongs to. To formulate this, if there are C classes in the dataset: $\{c_1, c_2, ..., c_i, ..., c_C\}$, and the number of samples in the class that the sample x belongs to is denoted by $|c_x|$, then:

$$\alpha_x = \frac{\frac{1}{|c_x|}}{\sum_{i=1}^{C} \frac{1}{|c_i|}} \tag{5.3}$$

where $|c_i|$ is the number of samples in the i^{th} class. α_x is the second factor of weighting sample x in contributing to the model's learning. The final weights, W^{t+1} are updated as in (5.4):

$$W^{t+1} = W^t - \eta \cdot \nabla \cdot \alpha_x \cdot \frac{L(y, f(x, w))}{\text{influence weight}(x; w)} \tag{5.4}$$

with W^t denoting the current weights and η, the learning rate.

5.3 Proposed Method

In this section, we propose a multi-modal approach for tackling the challenge of imbalanced document image classification, which combines both image-based and text-based information to effectively address class imbalance. As shown in Figure 5.1,

our approach consists of three main steps. For the visual stream, we employ a two-phase classification strategy using ResNet. During the first phase, the model is trained with cross-entropy loss; in the subsequent phase, we employ the IB loss function to mitigate the effects of imbalanced data.

Next, an OCR engine is utilized to extract textual content from document images, and BERT is employed for text classification. Lastly, the outputs of ResNet and BERT models are combined, and appropriate techniques are applied to mitigate class imbalance. Our method aims to improve both the accuracy and robustness of imbalanced document image classification tasks. The following sections detail each step, describing the methods used and the rationale behind our approach.

5.3.1 Visual stream

For classifying images based on their visual features, a robust CNN is needed. In the previous section, we looked at how various CNNs have been used to classify document images. In our study, we experimented with multiple CNNs, including well-known architectures like VGGNet [127], which is commonly used in image classification, ResNet [25], known for its deep layers and skip connections, and EfficientNet [128], a more recent model that has shown promise in various tasks. Although each of these networks had their own strengths, in our case their performance were roughly the same. To maintain clarity and focus in our discussions, we have decided to primarily discuss the results from ResNet.

A two-stage approach is implemented to address the imbalanced task, as shown in Figure 5.1. In the first phase, the model is trained for a specified number of epochs, denoted as T_1, utilizing the cross-entropy loss function to establish a baseline performance. In the second phase, which covers the remaining $T - T_1$ epochs, the training process continues and incorporates the IB loss. The IB loss is a specialized loss function designed for imbalanced data. It evaluates the relative influence of each training sample and applies class-wise re-weighting to mitigate the impact of influential instances. This approach leads to a more balanced and robust training process, ultimately improving the model's performance on the imbalanced task.

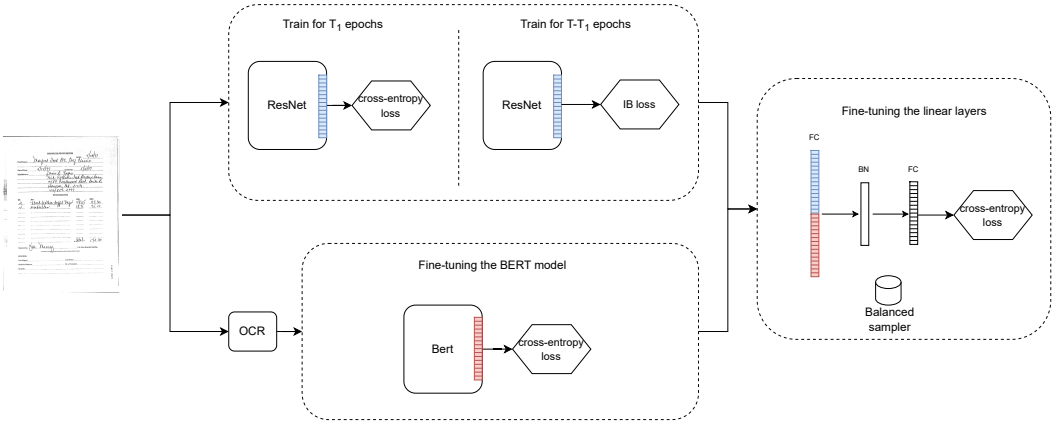
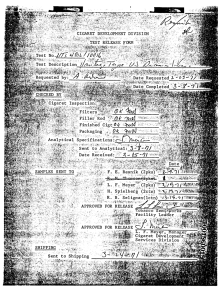


Fig. 5.1: Overview of the Proposed Multi-Modal Approach. It involves three primary parts, including the utilization of a two-phase classification strategy using ResNet, the incorporation of advanced OCR techniques for text classification using BERT, and the fusion of outputs from ResNet and BERT models.



CIGARET DEVELOPMENT DIVISION TEST RELEASE FORM Test No. VA 91001 Test Description Requested by: CHECKED BY Date Requested Date Completed Cigaret Inspection: Filters Filler Rod Packaging OK A44 ok mms4 OK h Finished Cigtok hwt4 Analytical Specifications Sent to Analytical: - Date Received: 2: 4s-n Date yc SAMPLES SENT TO F. E. Resnik (2pks) 2-14. L. F. Meyer (2pks) H. Spielberg (2ctn) R. B. Seligman(lctn) ke * APPROVED FOR RELEASE Tobacco Semi SYIOM- Facility Leader APPROVED FOR RELEASE T F. Meyer, Manager Cigaret D00e9velopment Services Division SHIPPING 0098100 Sent to Shipping

Auer Completed 3 1-7" 1m “Filler Rod “Ot “mal Doman Finished cigt 06 Seid 3 2 Packaging -. O aK of oma Sent ‘te “analytical?” 34-2 832 ie -Q or femi<plorks Facility Ledder : jeygr, Manager, Cigaret Dévélopment Services Division I

(a) Sample image **(b)** docTR **(c)** Tesseract

Fig. 5.2: Visualization the OCR results of docTR and tesseract engines.

5.3.2 Textual stream

The BERT model is a highly effective language model that has been pre-trained on massive textual data. Due to its great encoder representation for text, it is widely used for text classification tasks. It is common practice to achieve high performance by fine-tuning a pre-trained BERT on a small dataset with only a few epochs.

To begin, we first extract text from the document images using docTR - an advanced OCR tool developed by Mindee [129]. Figure 5.2 showcases the superiority of docTR over tesseract by presenting two sample dataset images along with their respective OCR outputs from docTR and tesseract. The extracted text is then fed into the BERT model for training. The input text is tokenized, padded and truncated to a fixed length to ensure consistency. For our classification task, we simply add a classification head on top of BERT’s pooled output. After training for a few epochs using cross-entropy loss, the model adapts well to our specific task. The output of

this textual stream model can be used independently or combined with the visual stream output, which we discuss next.

5.3.3 Fusion network

The core of our multi-modal document classification framework lies in the fusion of image and text networks, with the goal of leveraging the combined strengths of visual and textual data. This fusion combines the discriminative capabilities of visual features extracted by ResNet with the contextual understanding derived from textual content using BERT, enriching the representation of document images to improve classification accuracy.

After classifying images and text separately, each stream produces a 16-dimensional vector representing class probabilities. These vectors are combined into a single 32-dimensional vector. During this process, it is crucial to mention that all layers before the fully connected and fusion layer are frozen. This means that the pre-trained features are preserved and not modified during this phase.

Due to the inherent differences between visual and textual information, the fused vector may contain noise. To mitigate this, we utilize batch normalization on the concatenated vector to normalize its elements, enhancing generalization and learning efficiency across classes. Following batch normalization, a normalized vector is obtained and passed through a final classification layer. The result is an output vector that merges information from both modalities, capturing precise classification probabilities for each class.

Moreover, the fusion process integrates a class-balanced sampler technique to ensure equal representation of all classes by oversampling underrepresented classes. By deliberately oversampling the underrepresented classes, the model gains better familiarity with less common classes during training. This proactive step helps prevent bias towards majority classes, especially crucial as the final connected layers are prone to overfitting. In essence, this strategy results in enhanced overall performance in classification assignments.

5.3.4 Implementation details

The implementation and training of all models in the paper were conducted using PyTorch [130]. The official implementations of ResNet-50 and BERT were used for

image and text classification tasks, respectively. These models are trained on a single NVIDIA GTX 3090.

For the visual model, the input image was resized to 256x256. The training was conducted over 16 epochs using the cross-entropy loss. Following that, an additional 8 epochs were dedicated to training with the IB loss, resulting in a total training span of 24 epochs. The initial learning rate was set at 4e-4, with a batch size of 64 for data processing.

For the text model, we selected the base pre-trained model of BERT. The model was fine-tuned for four epochs, starting with a initial learning rate of 2e-5 with decay. A batch size of 32 was used for this training process. These exact hyper-parameters were also employed for training the fusion model.

Throughout all training processes, the AdamW optimizer [131] was employed. To dynamically adjust the learning rate, the cosine scheduler with warmup was applied. This scheduler managed the decrease of the learning rate by utilizing the values of the cosine function between the initial learning rate set in the optimizer and 0. The scheduling took place after a warmup period, during which the learning rate gradually increased from 0 to the initial rate defined in the optimizer.

5.4 Experiments and Results

5.4.1 Dataset

The RVL-CDIP dataset is widely recognized as the benchmark standard for evaluating document classification model performance. It consists of a collection of 400,000 grayscale images across 16 different classes. These classes represent various document types, making it an invaluable resource for training machine learning models. Notably, Figure 5.3 showcases a selection of sample images from this dataset. Originally, the dataset is balanced, with each class in the training set containing 20,000 images and each class in the validation and test sets containing 2,500 images. However, to examine the effects of our contribution, we have introduced a protocol to create an unbalanced version of the RVL-CDIP dataset.

In a long-tailed distribution, the sample sizes across classes exhibit an exponential decay pattern. This decay is characterized by the parameter ρ, which represents the ratio between the most and least frequent classes. To align with common practices in the literature, we have set ρ to 0.01. Consequently, in our modified dataset, the

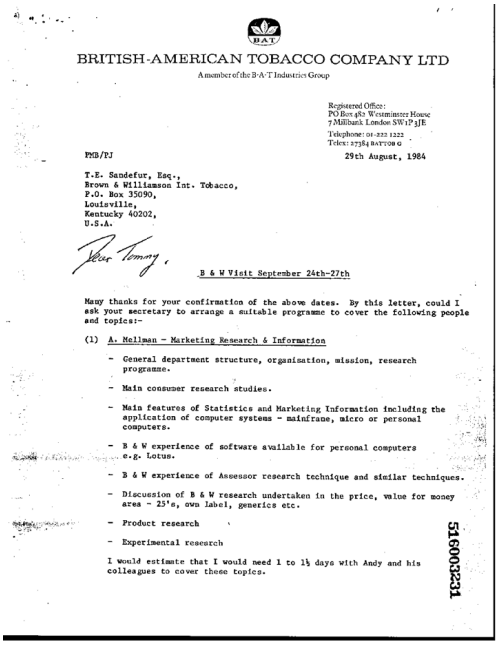
first class contains 20,000 samples, while the last class is represented by only 200 samples. Figure 5.4 provides a visual representation of the class distribution in our imbalanced train set. In order to ensure a fair comparison, we have maintained the original test set without any modifications.

To evaluate our approach, multiple experiments were conducted and the results were documented. We compare our method against two SOTA methods, namely LayoutLMv2 and DocFormer, as well as a baseline model, to highlight its performance enhancements in handling imbalanced datasets.

- **LayoutLMv2**: An advanced model that combines textual and layout information for document understanding, representing a significant benchmark in the field [132].
- **DocFormer**: A novel transformer architecture that integrating visual and spatial features throughout the model using residual connections [133].
- **ResNet+CE**: Utilizes the ResNet architecture with cross-entropy loss, setting the baseline for performance comparison.
- **ResNet+CE+IB**: Enhances the baseline by incorporating IB loss alongside cross-entropy loss.
- **ResNet+CE+IB+Focal**: Further extends the previous setup by adding Focal loss to address class imbalance more effectively.
- **BERT**: Employs the BERT model to focus on textual data analysis, diverging from visual features.
- **Proposed Model**: Our approach that integrates the strengths of ResNet trained with CE+IB+Focal losses and BERT, aiming to leverage both visual and textual data for improved classification performance.

5.4.2 Evaluation Metrics

Evaluation metrics are essential in both assessing classification performance and guiding model development. While accuracy is commonly used, it becomes problematic when dealing with class imbalance issues. In such cases, a high accuracy score may not accurately reflect the classifier's ability to identify rare classes. For instance, if only 1 percent of training data belongs to a rare class, an extremely simple strategy can achieve high accuracy by always predicting the majority class. This highlights



BRITISH-AMERICAN TOBACCO COMPANY LTD

A member of the B·A·T Industries Group

Registered Office:
PO Box 483 Westminster House
7 Millbank London SW1P 3JE
Telephone: 01-222 1222
Telex: 27384 BATTOR G

FMB/PJ

29th August, 1984

T.E. Sandefur, Esq.,
Brown & Williamson Int. Tobacco,
P.O. Box 35090,
Louisville,
Kentucky 40202,
U.S.A.

Dear Tommy,

B & W Visit September 24th-27th

Many thanks for your confirmation of the above dates. By this letter, could I ask your secretary to arrange a suitable programme to cover the following people and topics:-

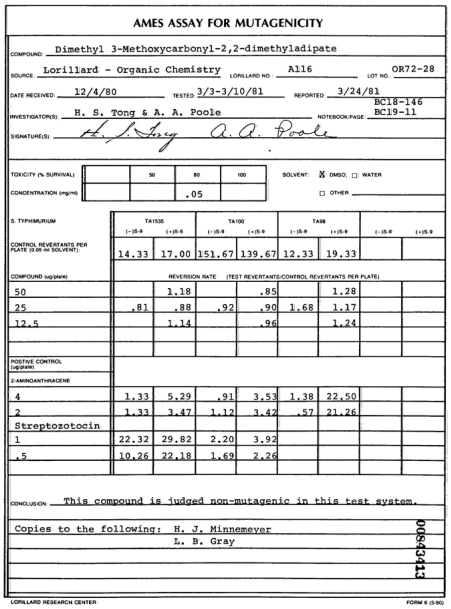
(1) A. Mellman - Marketing Research & Information

- General department structure, organisation, mission, research programme.
- Main consumer research studies.
- Main features of Statistics and Marketing Information including the application of computer systems - mainframe, micro or personal computers.
- B & W experience of software available for personal computers e.g. Lotus.
- B & W experience of Assessor research technique and similar techniques.
- Discussion of B & W research undertaken in the price, value for money area - 25's, own label, generics etc.
- Product research
- Experimental research

I would estimate that I would need 1 to 1½ days with Andy and his colleagues to cover these topics.

516003231

(a) letter



AMES ASSAY FOR MUTAGENICITY

COMPOUND: Dimethyl 3-Methoxycarbonyl-2,2-dimethyladipate

SOURCE: Lorillard - Organic Chemistry LORILLARD NO.: A116 LOT NO.: OR72-28

DATE RECEIVED: 12/4/80 TESTED: 3/3-3/10/81 REPORTED: 3/24/81

(INVESTIGATOR(S): H. S. Tong & A. A. Poole NOTEBOOK/PAGE: BC18-146 BC19-11

SIGNATURE(S): H. S. Tong A. A. Poole

TOXICITY (% SURVIVAL): 50 80 100

CONCENTRATION (mg/ml): .05

SOLVENT: ☒ DMSO; ☐ WATER ☐ OTHER

S. TYPHIMURIUM

	TA1535 (-)S-9	TA1535 (+)S-9	TA100 (-)S-9	TA100 (+)S-9	TA98 (-)S-9	TA98 (+)S-9	(-)S-9	(+)S-9
CONTROL REVERTANTS PER PLATE (0.05 ml SOLVENT):	14.33	17.00	151.67	139.67	12.33	19.33		
COMPOUND (ug/plate)	REVERSION RATE (TEST REVERTANTS/CONTROL REVERTANTS PER PLATE)							
50		1.18		.85		1.28		
25	.81	.88	.92	.90	1.68	1.17		
12.5		1.14		.96		1.24		
POSITIVE CONTROL (ug/plate)								
2-AMINOANTHRACENE								
4	1.33	5.29	.91	3.53	1.38	22.50		
2	1.33	3.47	1.12	3.42	.57	21.26		
Streptozotocin								
1	22.32	29.82	2.20	3.92				
.5	10.26	22.18	1.69	2.26				

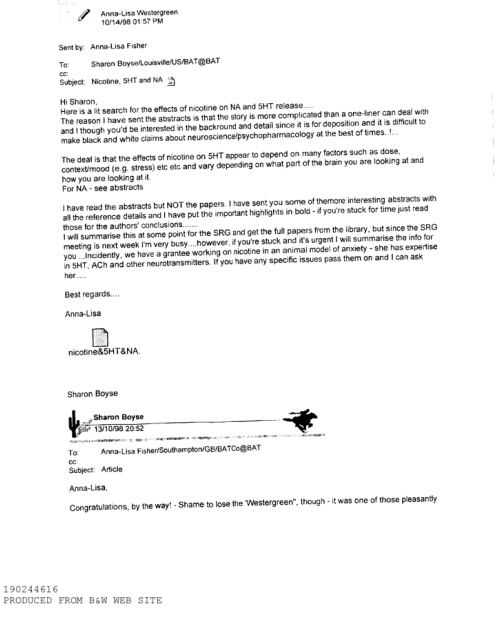
CONCLUSION: This compound is judged non-mutagenic in this test system.

Copies to the following: H. J. Minnemeyer
L. B. Gray

00643413

LORILLARD RESEARCH CENTER FORM 6 (5-80)

(b) form



Anna-Lisa Westergreen
10/14/98 01:57 PM

Sent by: Anna-Lisa Fisher

To: Sharon Boyse/Louisville/US/BAT@BAT
cc:
Subject: Nicotine, 5HT and NA

Hi Sharon,
Here is a lit search for the effects of nicotine on NA and 5HT release.....
The reason I have sent the abstracts is that the story is more complicated than a one-liner can deal with and I though you'd be interested in the backround and detail since it is for deposition and it is difficult to make black and white claims about neuroscience/psychopharmacology at the best of times..!...

The deal is that the effects of nicotine on 5HT appear to depend on many factors such as dose, context/mood (e.g. stress) etc etc and vary depending on what part of the brain you are looking at and how you are looking at it.
For NA - see abstracts

I have read the abstracts but NOT the papers. I have sent you some of themore interesting abstracts with all the reference details and I have put the important highlights in bold - if you're stuck for time just read those for the authors' conclusions.....
I will summarise this at some point for the SRG and get the full papers from the library, but since the SRG meeting is next week I'm very busy....however, if you're stuck and it's urgent I will summarise the info for you ...Incidently, we have a grantee working on nicotine in an animal model of anxiety - she has expertise in 5HT, ACh and other neurotransmitters. If you have any specific issues pass them on and I can ask her.....

Best regards....

Anna-Lisa

nicotine&5HT&NA.

Sharon Boyse

Sharon Boyse
13/10/98 20:52

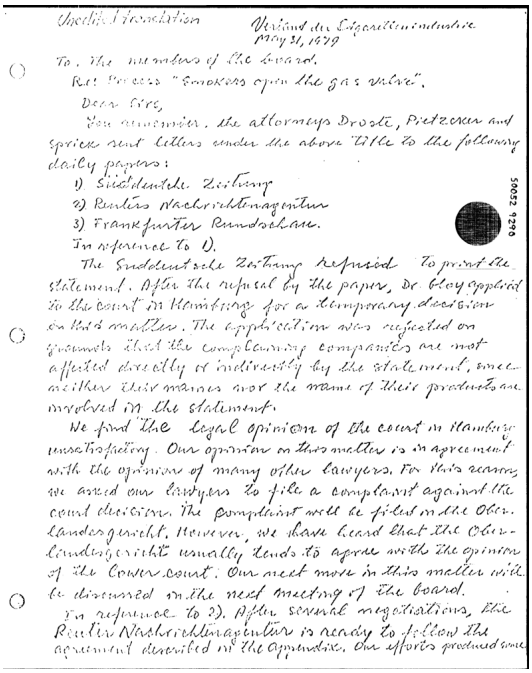
To: Anna-Lisa Fisher/Southampton/GB/BATCo@BAT
cc:
Subject: Article

Anna-Lisa,

Congratulations, by the way! - Shame to lose the 'Westergreen", though - it was one of those pleasantly

190244616
PRODUCED FROM B&W WEB SITE

(c) email



Verband der Cigarettenindustrie
May 31, 1979

To: The members of the board.

Re: Press "Smokers open the gas valve".

Dear Sirs,

You remember, the attorneys Droste, Pretzecker and Spricks sent letters under the above title to the following daily papers:

1) Süddeutsche Zeitung
2) Reuters Nachrichtenagentur
3) Frankfurter Rundschau.

In reference to 1).

The Süddeutsche Zeitung refused to print the statement. After the refusal by the paper, Dr. Bley applied to the court in Hamburg for a temporary decision in that matter. The application was rejected on grounds that the complaining companies are not affected directly or indirectly by the statement, since neither their names nor the name of their products are involved in the statement.

We find the legal opinion of the court in Hamburg unsatisfying. Our opinion on this matter is in agreement with the opinions of many other lawyers. For this reason we asked our lawyers to file a complaint against the court decision. The complaint will be filed in the Oberlandesgericht. However, we have heard that the Oberlandesgericht usually tends to agree with the opinion of the lower court. Our next move in this matter will be discussed in the next meeting of the board.

In reference to 2). After several negotiations, the Reuter Nachrichtenagentur is ready to follow the agreement described in the Appendix. Our efforts produced some

59052 9290

(d) handwritten

Fig. 5.3: Sample images from RVL-CDIP dataset.

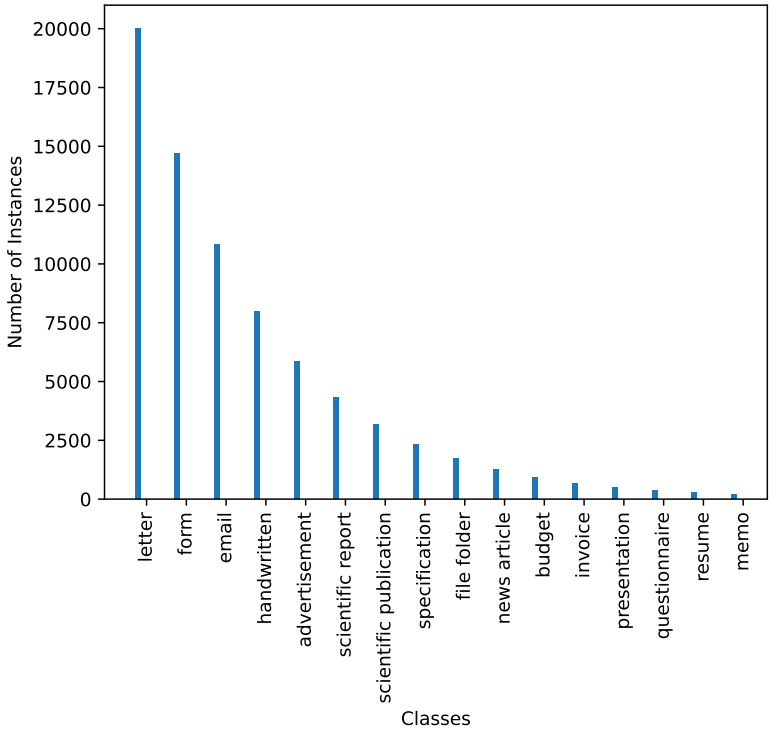
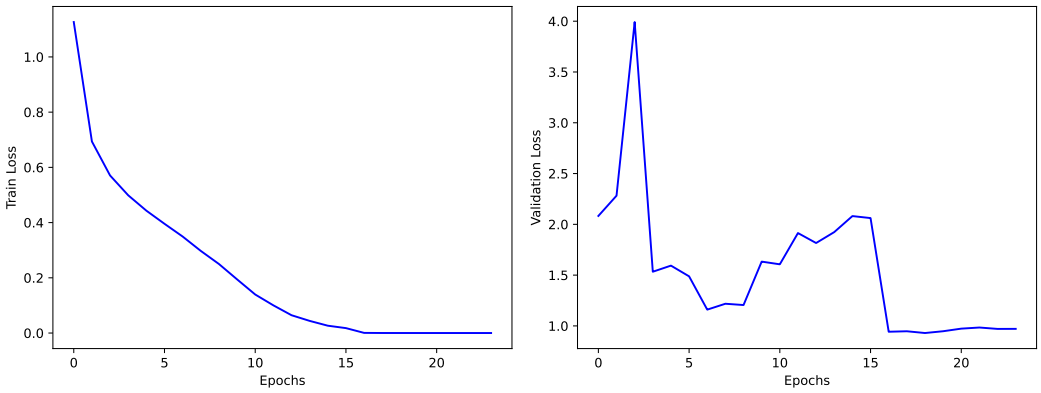


Fig. 5.4: Class frequency distribution in imbalanced training set.



(a) training loss **(b)** validation loss

Fig. 5.5: Loss curve.

the limitations of using accuracy as the sole evaluation metric in certain applications where identifying rare cases is crucial.

Accuracy

Accuracy is the ratio of correctly predicted instances to the total instances. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.5}$$

where:

- TP (True Positive): Correctly predicted positive instances
- TN (True Negative): Correctly predicted negative instances
- FP (False Positive): Incorrectly predicted positive instances
- FN (False Negative): Incorrectly predicted negative instances

F1 Score

The F1 score is the harmonic mean of precision and recall, providing a balance between them. It is defined as:

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.6}$$

where:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.8}$$

Confusion Matrix

The confusion matrix is a table that summarizes the performance of a classification model by displaying the actual vs. predicted classifications. It is structured as follows:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Tab. 5.1: Confusion Matrix

5.4.3 Results

Table 5.2 presents the results, where we have 16 classes, and multiple columns correspond to different experimented approaches under each metric. The analysis of accuracy and F1-scores reveals a trend of decreasing performance as the number of instances per class decreases. Notably, some majority classes achieve high accuracy but lower F1-scores. For instance, the 'form' class exhibits this phenomenon, with high accuracy but a lower F1-score, indicating a significant number of mislabeled samples despite the majority being correctly predicted. This decline in performance is consistent with the challenges associated with class imbalance, where models tend to underperform on minority classes. Specifically, LayoutLMv2 demonstrates good performance across most classes but struggles with less represented classes, while DocFormer displays the weakest performance in these minority classes.

Comparing the baseline ResNet+CE to the ResNet+CE+IB demonstrates noticeable improvements. Furthermore, incorporating Focal loss with the IB loss leads to additional gains in many cases, showing the complementary effects of these strategies. The impact of using IB loss is also visible in the loss curves shown in Figure 5.5. Initially, while the training loss decreased, the validation loss showed signs of overfitting. However, after the sixteenth epoch, when training continued with IB loss, the validation loss noticeably dropped, demonstrating the effectiveness of this loss.

Markedly, the BERT model shows greater resilience, potentially attributed to its ability to leverage distinct textual information for accurate classification. For instance, the 'resume' class, despite limited samples, showcases BERT's capability to identify it through unique textual cues.

In the baseline approaches, the majority classes display high accuracy but low F1-score. This pattern indicates that samples from potentially minority classes are being misclassified as belonging to the more populated majority classes, which is a common issue when dealing with imbalanced datasets.

The proposed method, however, has been successful in maintaining a balanced performance across both the majority and minority classes.

Tab. 5.2: Comparative Analysis of Accuracy and F1-Score Across Different Models and Classes. The highest values for each metric and class highlighted in bold.

Class	Accuracy							F1-Score						
	LayoutLMv2	DocFormer	ResNet+ CE	ResNet+ CE+IB	ResNet+ CE+IB+Focal	BERT	Proposed	LayoutLMv2	DocFormer	ResNet+ CE	ResNet+ CE+IB	ResNet+ CE+IB+Focal	BERT	Proposed
letter	**94.5**	89.4	93.7	92.2	90.8	94.2	92.0	64.0	54.2	64.9	72.0	74.8	72.5	**86.9**
form	87.3	84.8	89.0	86.9	87.2	**90.6**	88.4	63.5	47.9	57.1	60.9	62.7	70.3	**80.6**
email	98.2	**98.6**	98.1	98.0	98.2	97.9	97.7	95.4	84.2	96.2	96.2	96.9	96.2	**97.4**
handwritten	92.7	87.2	95.3	95.3	**95.5**	93.8	95.0	**93.9**	86.2	91.4	91.0	91.2	88.2	89.9
advertisement	**93**	81.8	91.7	90.8	91.1	85.9	88.8	84.9	75.2	83.0	83.0	84.3	81.9	**87.9**
scientific report	78.3	62.2	75.3	73.8	74.8	**86.0**	85.1	64.1	64.8	60.4	64.5	64.0	77.7	**81.5**
scientific publication	86.3	85.3	88.4	88.5	87.8	88.3	**90.9**	88.1	81.8	87.1	87.3	87.5	89.4	**91.2**
specification	83	80.0	81.7	82.7	83.4	89.3	**90.1**	87.5	84.5	85.0	86.5	85.3	92.8	**92.8**
file folder	90.6	84.1	90.1	92.6	90.9	84.2	**95.3**	**91.2**	81.3	89.0	88.4	88.7	80.9	86.1
news article	80.6	66.2	76.0	78.4	79.6	78.4	**82.8**	84.5	69.8	82.4	82.3	83.3	84.5	**87.0**
budget	72.1	51.5	58.7	61.5	64.7	76.3	**81.7**	75.5	61.6	69.1	70.7	71.0	81.6	**84.0**
invoice	62.7	27.1	50.1	57.4	62.7	74.6	**83.3**	74.7	41.9	65.0	69.8	73.1	83.1	**87.1**
presentation	49.9	36.1	44.7	52.1	50.9	56.9	**69.6**	61.8	49.2	58.2	62.4	61.9	68.6	**76.6**
questionnaire	48.7	32.7	37.7	46.2	49.2	63.0	**80.0**	62.9	47.5	53.7	61.0	63.4	76.9	**86.3**
resume	71.3	84.0	71.8	78.7	79.7	95.9	**96.6**	83.0	90.6	83.0	86.3	86.6	97.8	**98.0**
memo	39.6	6.3	40.6	58.7	62.8	56.8	**84.1**	56.3	11.9	57.4	72.3	75.8	72.0	**88.5**
average	76.9	66.2	74.0	77.2	78.1	82.0	**87.6**	77.0	64.6	74.0	77.1	78.1	82.1	**87.6**

The fusion method consistently outperforms others in handling class imbalance challenges, as evidenced by the results. It significantly boosts F1-scores for classes with limited sample sizes, such as ‘resume‘, ‘questionnaire‘, and ‘memo‘.

Furthermore, the confusion matrix, represented as $CM(i, j)$, provides a visual summary of multi-class classification performance (Figure 5.6). Each column corresponds to an actual class, while each row represents a predicted class. The diagonal elements of the confusion matrix indicate accurate classification predictions where $i = j$.

Taking a closer look at the confusion matrix, we can identify areas where the model faces difficulty distinguishing between very similar classes. For example, the ‘memo‘ and ‘letter‘ classes are challenging due to their resemblance, particularly since the ‘memo‘ class has only 200 samples. Similarly, the ‘budget‘ and ‘invoice‘ classes have similarities that make telling them apart tricky. Additionally, some samples are ambiguous and could belong to more than one label. For instance, the class “questionnaire” includes handwritten forms that are filled out.

Considering the provided analysis, the proposed approach has effectively preserved accuracy in the majority classes, which naturally demonstrate high accuracy, while also substantially improving accuracy in the minority classes. The results also prove the effectiveness of the fusion of the image and text modalities.

5.5 Conclusion

This chapter addressed the significant challenge of class imbalance in document image classification, where some classes have far fewer examples than others. To tackle this issue, a novel multi-modal approach was developed, combining both visual and textual information from document images.

The proposed method consisted of two streams: a visual stream using a ResNet architecture with Influence-Balanced loss, and a textual stream using an OCR engine and the BERT model. The outputs of these two streams were fused and normalized using batch normalization to reduce noise and improve generalization.

The approach was evaluated on a modified version of the RVL-CDIP dataset, designed to simulate real-world imbalanced scenarios. The results showed that the proposed method outperformed state-of-the-art methods, including LayoutLMv2 and DocFormer, as well as baseline models.

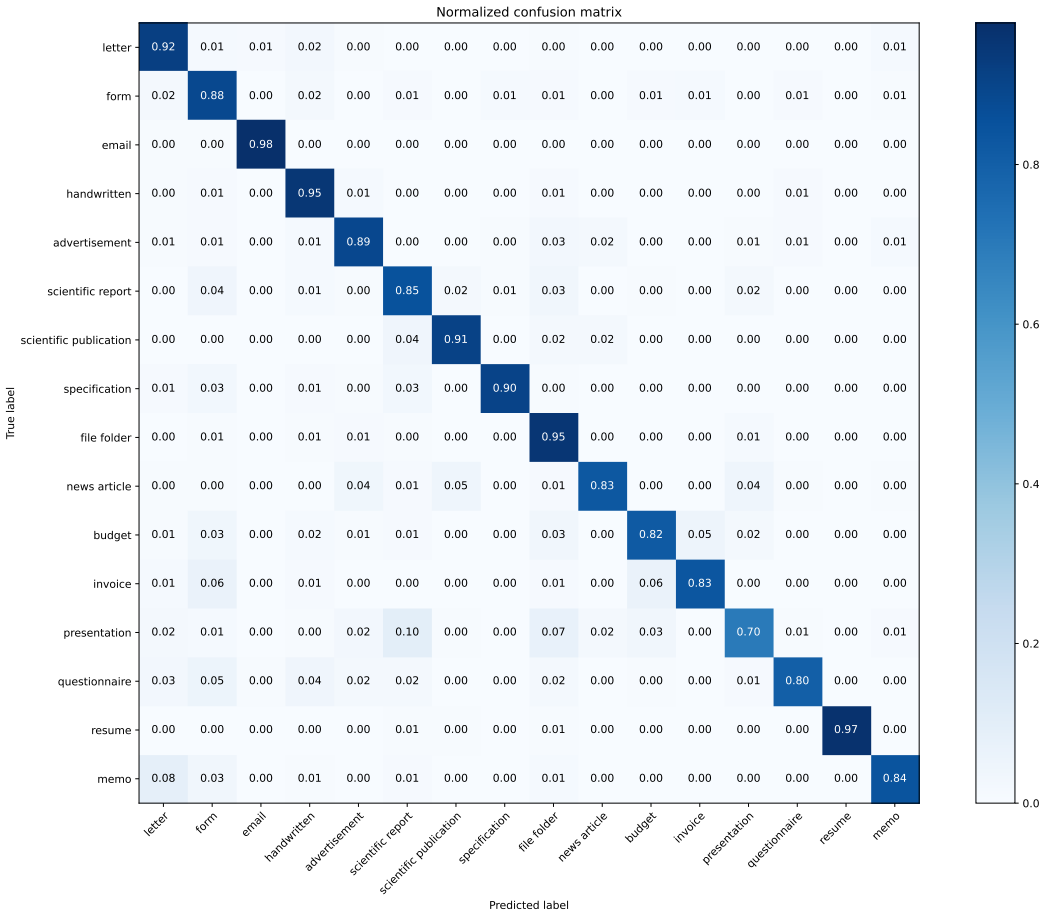


Fig. 5.6: Confusion matrix of the proposed method.

The experiments demonstrated a significant performance improvement, with an overall accuracy boost of 13 percent and a more than 40 percent improvement in certain minority classes. The results highlighted the importance of combining both visual and textual features and employing specialized techniques, like IB loss, to address the negative impact of imbalanced datasets.

In summary, this research introduced a novel multi-modal approach that effectively addresses the challenges of class imbalance in document classification. By combining ResNet with IB loss, BERT, and a fusion strategy, the method significantly improved overall accuracy and performance on minority classes. This work provides a solid foundation for future research in this area and underscores the importance of using specialized techniques for imbalanced datasets.

6 Structural Information Extraction Using LLMs

> *Don't hesitate to ask the universe for more; you probably won't get it anyway.*
>
> — **Dave Tarnowski**

6.1 Introduction

Document understanding aims to interpret and extract meaningful information from documents, which is made challenging by the numerous variations in document layouts, such as invoices, tax forms, and other structured formats. These documents often have complex elements such as tables and key-value pairs, necessitating sophisticated methods for accurate information extraction. The field evolved from heuristic methods [134] to specialized deep neural network techniques [135] and, more recently, to the application of LLMs [136].

Natural language processing has been transformed by the advent of LLMs, which have outperformed earlier SOTA methods in tasks such as text understanding, generation, and information extraction. Models like GPT [137] and Llama [138] have demonstrated remarkable capabilities thanks to their huge scale (often consisting of hundreds of billions of parameters) and extensive training on a wide range of datasets. These models can perform tasks without the need for task-specific training data due to their exceptional performance in zero-shot and few-shot learning. However, applying LLMs to visually dense, structured documents remains difficult due to a lack of layout information in the text, which is required for accurate information extraction.

This chapter explores the use of LLMs for structural information extraction, with a focus on fine-tuning techniques and the integration of layout information via HTML representations. We begin by reviewing existing methods for documenting understanding, followed by a thorough discussion of LLMs and their fine-tuning process. Finally, we present our proposed method, which uses HTML representations and

tailored instruction prompts to improve LLM performance on structured document tasks.

6.2 Background

The development of models that incorporate text, layout, and visual information has resulted in significant advances in document understanding. The LayoutLM series [139, 132, 135] takes a major step forward by combining language models with spatial and visual contexts. LayoutLM introduced 2-D positional and image embeddings for tokens, while LayoutLMv2 improved visual data integration with a multimodal transformer architecture. LayoutLMv3 simplified the structure by using Vision Transformers [140] and focusing on tasks like masked language modeling and word-patch alignment.

FormNet [141] employed innovative techniques like rich attention and super-tokens to capture structural details in documents. This procedure computes attention scores based on spatial relationships between tokens, whereas super-tokens incorporate embeddings from neighboring tokens via graph convolutions. These approaches have produced impressive results when dealing with structured documents.

Donut [142] proposed an end-to-end encoder-decoder model that bypasses OCR by directly mapping raw input images to desired outputs. This approach simplifies the pipeline but may struggle with documents requiring precise layout understanding.

Recent work by Perot et al. [143] introduced LMDX, a method for extracting information from semi-structured documents using LLMs. LMDX uses text position encoding and a grounding mechanism in five stages: OCR, chunking, generating prompts, LLM inference, and decoding. This pipeline successfully identifies and locates entities in documents, demonstrating the power of LLMs for document understanding.

Despite these advancements, there are still challenges in using LLMs for structured document tasks, particularly due to a lack of layout information in text. Our work addresses this limitation by introducing a machine-friendly HTML representation that preserves both textual content and spatial layout, enabling more accurate and precise analysis.

6.2.1 Large Language Models

Large language models are a class of deep learning models designed to understand and generate human-like text. They are typically based on the Transformer architecture [144], which is a groundbreaking deep learning model that powers from self-attention mechanisms. LLMs are trained on large text corpora to predict the next words in a sequence, allowing them to identify complex linguistic patterns and semantic relationships, improving their language proficiency. Modern models, such as GPT [137] and Llama [138], have enormous size, often comprising hundreds of billions of parameters. can generalize across tasks without the need for task-specific training data.

The Transformer architecture consists primarily of two components: an encoder and a decoder. The encoder processes the input sequence to produce rich, continuous representations, and the decoder generates the output sequence using these encoder representations and previously generated tokens. The versatility of the Transformer architecture allows for a wide range of configurations, each tailored to a specific task. The encoder-decoder configuration is used for sequence-to-sequence tasks like machine translation when the input and output sequences differ. Encoder-only configurations, on the other hand, are adequate for tasks that require only the comprehension of an input sequence, such as text classification or sentiment analysis. The decoder-only configuration is used for generative tasks such as text generation or language modeling.

Figure 6.1 illustrates the original Transformer architecture as described in [144]. The key components of the Transformer architecture include the following:

- **Multi-head self-attention**: enables the model to focus on different parts of the input sequence simultaneously, allowing for the capture of complex relationships between tokens.
- **Positional encodings**: provide information about the order of tokens in the input sequence, which is essential for understanding the context and structure of the input data.
- **Feed-forward networks**: apply non-linear transformations to the outputs of the self-attention mechanism, enabling the model to refine its representations and generate more accurate outputs.
- **Layer normalization**: stabilizes the training process by normalizing the activations of each layer, which helps to prevent vanishing or exploding gradients.

- **Residual connections**: enable the model to learn much deeper representations than would be possible with a standard feed-forward architecture, by allowing the model to preserve and reuse previously learned features.

At the heart of LLMs lies the self-attention mechanism, which computes attention scores between all pairs of tokens in a sequence. Given an input sequence of token embeddings $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, the self-attention mechanism computes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{1}$$

where:

- $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_V$ are the **query**, **key**, and **value** matrices, respectively.

- $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learned weight matrices.

- d_k is the dimensionality of the key vectors, used to scale the dot product.

The output of the self-attention layer is a weighted sum of the value vectors, where the weights are determined by the compatibility between queries and keys. This allows the model to focus on relevant parts of the input sequence when generating text or making predictions.

LLMs are not limited to just generating text; they have the potential to transform how we understand documents. Their applications are diverse, including tasks such as extracting information, conducting semantic searches, and summarizing documents [136]. The few-shot learning capability of LLMs is particularly useful in domains where labeled data is scarce or expensive to obtain [145]. By providing a limited number of examples, LLMs can adapt to specific tasks, achieving SOTA performance without extensive fine-tuning. However, their effectiveness in document understanding tasks depends heavily on how input data is prepared and structured.

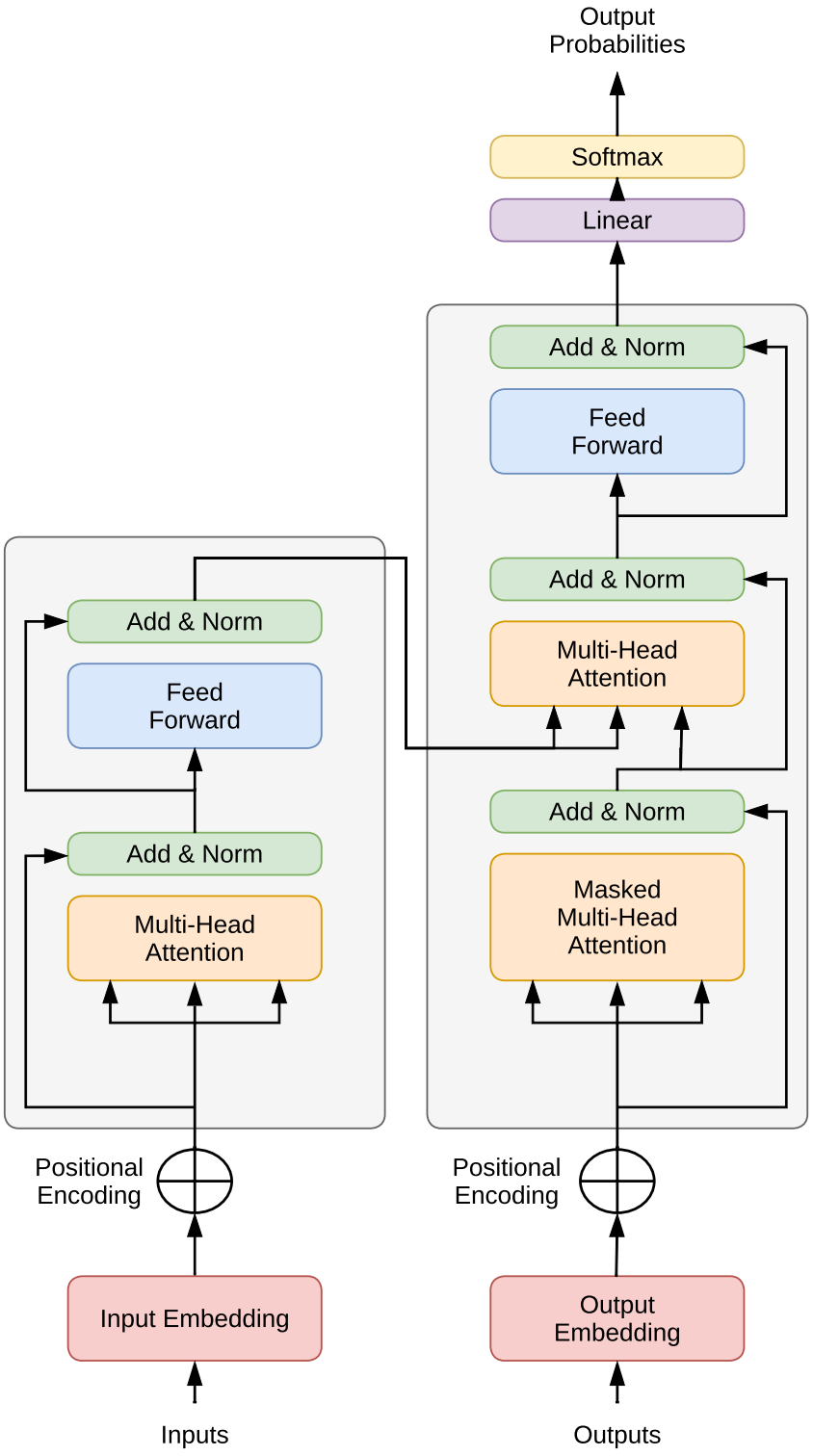


Fig. 6.1: Visual representation of the Transformer architecture, highlighting the encoder-decoder structure, self-attention and feed-forward layers, and positional encodings.

6.3 Methodology

6.3.1 Fine-Tuning Large Language Models

Fine-tuning is a supervised learning method that improves an LLM's ability to perform specific tasks by updating its weights with labeled datasets. Unlike pre-training, which involves training on massive amounts of unstructured textual data, fine-tuning focuses on task-specific adaptations. Instruction fine-tuning, in particular, has emerged as a powerful strategy for improving model performance on a wide range of problems. This method aligns the model's behavior with human-like instructions, allowing it to generalize more effectively for multiple tasks and domains.

Instruction fine-tuning involves training the model on datasets built up of prompt-completion pairs, each of which contains an instruction and the desired outcome. To improve a model's summarization ability, for example, the dataset could include prompts such as "Summarize the following text", along with appropriate completions. This method teaches the model how to generate responses that are consistent with the instructions provided. The ability of instruction fine-tuning to bridge the gap between the model's pre-trained knowledge and the specific requirements of downstream tasks is critical to its success.

During fine-tuning, the model generates completions for prompts from the training dataset and compares them to the labeled responses. The loss, calculated with the cross-entropy function, is used to update the model's weights via backpropagation. This process is repeated over several epochs to improve the model's performance on the intended task. The selection of hyperparameters, such as learning rate, batch size, and epoch count, is critical in determining the efficacy of fine-tuning. A precise learning rate, for example, ensures that the model converges to the best solution without overfitting or underfitting the data. A validation dataset is used to measure accuracy, while a test dataset is used to evaluate final performance. Fine-tuning creates a new version of the base model, also known as a "instructed model", that is more appropriate for the tasks for which it was fine-tuned. This process improves both task-specific performance and the model's ability to generalize to previously unseen but related tasks.

Our approach uses HTML representations to preserve both textual content and spatial layout, addressing the issue of missing layout information in text. We convert the OCR output of documents into HTML, which is used as input for the LLM. This input is accompanied by an instruction prompt that describes the task and the expected JSON output structure, which is based on the ground truth. By incorporating HTML,

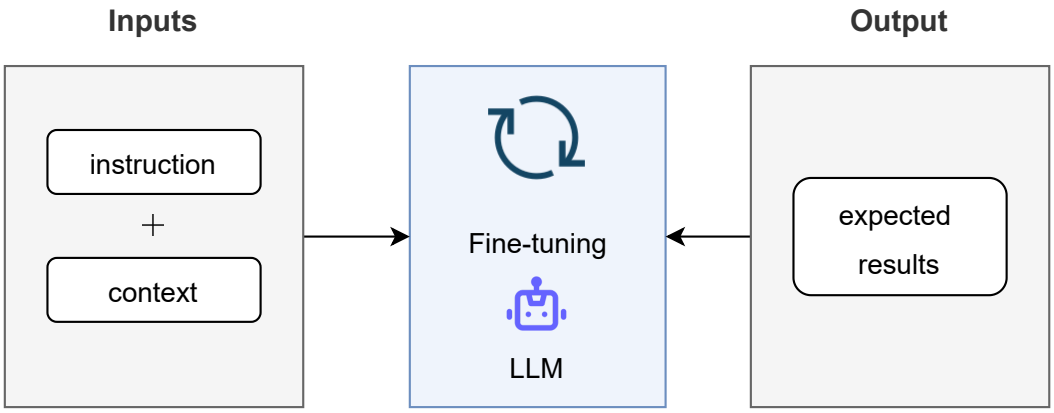


Fig. 6.2: An overall representation of our approach. Having prepared the inputs and expected outputs of the documents, the model is fine-tuned in a supervised manner.

we maintain the document's structural and semantic relationships, allowing the model to better understand and process complex layouts.

We fine-tune the LLM with tailored instruction prompts to help it better understand and process document content. The fine-tuned model generates JSON outputs containing key-value pairs for each page, making information extraction more accurate. Figure 6.2 shows the overall process, including data preparation, prompt generation, fine-tuning, and evaluation. This end-to-end pipeline ensures that the model is not only accurate, but also resistant to variations in document structure and content.

We use CodeLlama [146], a variant of Llama 2 [138] that has been fine-tuned for datasets containing code, markup languages, and natural language text. CodeLlama is specifically designed to handle long input contexts (up to 16,384 tokens) and performs well under detailed instructions, making it ideal for programming and data manipulation tasks. Its ability to handle structured inputs, such as HTML, makes it ideal for our document understanding tasks.

For fine-tuning, we use the Parameter-Efficient Fine-Tuning (PEFT) library [147] combined with a Low-Rank Adaptation (LoRA) configuration [148]. The PEFT library offers several options for fine-tuning a small number of extra parameters in large language models, significantly lowering computational costs. LoRA provides an efficient method for fine-tuning LLMs by introducing smaller low-rank matrices into each layer rather than directly modifying the original weight matrices. This method involves freezing the pre-trained model weights and injecting trainable

low-rank decomposition matrices into the model architecture. Only these smaller matrices are updated during training, which significantly reduces the number of trainable parameters and memory requirements while maintaining competitive performance.

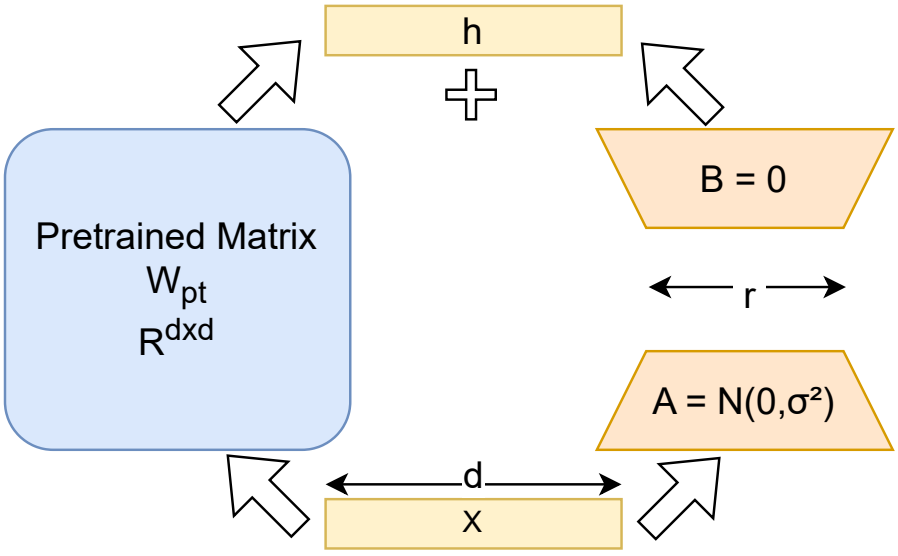


Fig. 6.3: Low-Rank Decomposition for Efficient Neural Network Fine-tuning. Schematic representation of low-rank adaptation in neural networks. The fine-tuned weights (W_{ft}) are decomposed into pre-trained weights (W_{pt}) plus a low-rank update (AB), where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$. The right panel illustrates the matrix dimensions and initialization strategy, where A is initialized from $\mathcal{N}(0, \sigma^2)$ and B from zero. This decomposition enables memory-efficient model adaptation while preserving the pre-trained weights.

Figure 6.3 illustrates the Low-Rank Adaptation framework. In this framework, the fine-tuned weights (W_{ft}) are represented as the sum of the original pre-trained weights (W_{pt}) and a weight update (ΔW). The key innovation lies in approximating this weight update using a low-rank decomposition, expressed as the product of two matrices A and B. While the original weight matrices exist in a high-dimensional space ($\mathbb{R}^{d\times d}$), the decomposition matrices A and B have reduced dimensions ($A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, respectively), where r is significantly smaller than d. This decomposition dramatically reduces the number of trainable parameters while maintaining model performance. The right panel of the figure provides additional implementation details, showing that matrix A is initialized with a normal distribution $\mathcal{N}(0, \sigma^2)$ and matrix B is initialized at zero, and illustrates how these matrices interact with the input (x) and hidden (h) dimensions of the network.

By utilizing LoRA, we achieve a balance between computational efficiency and model performance. This method allows us to fine-tune large models like CodeLlama on a single GPU, making it possible to adapt the model to multiple tasks without requiring extensive hardware resources.

6.3.2 HTML Representation

HTML is an ideal format for representing complex layout structures of documents. In the past, OCR engines such as Tesseract [149] have offered a specialized HTML representation in hOCR format [150]. Unlike plain text, HTML elements can capture how textual components spatially relate to one another within a rich formatting structure, which is particularly useful in scenarios that involve key-value inputs and require maintaining the relationships between words. As studied in [151], HTML serves as an interpretable structured medium for LLM.

In our application, we implemented a pre-processing step that transforms OCR output into HTML format. The details of this conversion is outlined in Algorithm 3. We use bounding box coordinates to arrange the text elements into a `<table>` layout, with `<tr>` rows and `<td>` cells, based on their relative positions. The algorithm then sorts and organizes these elements to create a coherent HTML structure that retains original spatial positioning relationships. Figure 6.4 shows a sample document and corresponding HTML encoding generated by this process.

Algorithm 3: Convert OCR results to HTML Table

Data: List of texts and bounding boxes

Result: HTML table

```
Function ConvertOCRtoHTML(List of texts and bounding boxes):
    foreach bounding box do
        Calculate row and column based on bounding box coordinates;
        Append (row, column, text) to data list;
    Sort data_list by row and then by column;
    Initialize table_html;
    foreach (row, column, text) in sorted_data do
        if row ≠ current_row then
            if current_row ≠ 0 then
                Add "</tr>" to table_html;
            Add "<tr>" to table_html;
            Update current_row and reset last_col;
        Calculate colspan based on column and last_col;
        if colspan > 0 then
            Add empty <td> with colspan to table_html;
        Add <td> with text content to table_html;
        Update last_col with current column;
    Add "</tr></table>" to table_html;
```

6.3.3 Prompt Generation

LLMs can be guided to perform specific tasks by using instruction prompts that clearly define the expected behavior. The Llama LLM uses two types of prompts: a system prompt and an instruction prompt. The system prompt sets the general tone and expectations for the interaction and is placed at the beginning. Meanwhile, the instruction prompt specifies the particular task or type of response expected from the model for that exchange.

We have designed the system prompt as follows:

> *"Below is an instruction that describes a task, paired with an input that provides further context. Your response is a JSON object that appropriately completes the request. The JSON must be between [JSON] and [/JSON] tags."*

KATZ TELEVISION GROUP

125 West 55th St
New York, NY 10019

Contract # 26790171 | **Changes as of:** 1/27/2020 at 11:42 AM | **Version:** Original Order

CPE: 23/26/383 | **Flight:** 2/17/20 - 2/23/20 | **Station:** KTAB | **Con Type:** POLITICAL/VOTE

Agency: DAVIS LENZ MEDIA
6060 N CENTRAL EXPRESSWAY
SUITE 560
DALLAS, TX 75206

Advertiser: Jon Francis Campaign | **Market:** Abilene | **Total $:** $5,135.00

Product: POLITICAL | **Office:** DALLAS | **Total Spots:** 24

Agency Order #: 9307478 | **Service:** Nielsen | **Total CPP:** $0.00

Buyer: Beth Davis, Haley | **Primary Demo:** | **Total GRP:**

Salesperson: ANDREA KRAUS
214-525-2617 | **Assistant:** JACOB BURNETT
214-525-2681 | **Traffic #:** 2357708

Separation: Sep:15

Comments: Separation: 15

#	Day/Time	DP	Program	Rate	Len	2/17 - 2/23							Total Spots	Total $	CPP*	GRP*
						2/17	2/18	2/19	2/20	2/21	2/22	2/23				
1	M-F 5:30a-7a		KTAB Daybreak	$145.00	30	1	1	1	1	1	0	0	5	$725.00	$0.00	0.0
2	M-F 7a-9a		CBS This Morning	$110.00	30	1	0	0	1	0	0	0	2	$220.00	$0.00	0.0
3	M-F 10a-11a		The Price is Right	$280.00	30	1	0	1	0	0	0	0	2	$560.00	$0.00	0.0
4	M-F 12n-12:30p		KTAB Noon News	$115.00	30	0	1	0	1	0	0	0	2	$230.00	$0.00	0.0
5	M-F 2p-3p		Lets Make A Deal	$105.00	30	1	0	1	0	0	0	0	2	$210.00	$0.00	0.0
6	M-F 4:30p-5p		Jeopardy	$115.00	30	0	0	1	0	0	0	0	1	$115.00	$0.00	0.0
7	M-F 5p-5:30p		KTAB 5P News	$190.00	30	1	1	0	1	0	0	0	3	$570.00	$0.00	0.0
8	M-F 6p-6:30p		KTAB 6 P News	$375.00	30	1	1	1	1	0	0	0	4	$1,500.00	$0.00	0.0
9	M-F 6:30p-7p		Wheel of Fortune	$450.00	30	0	1	0	1	0	0	0	2	$900.00	$0.00	0.0
10	Su 9:30a-10a		Face the Nation	$105.00	30	0	0	0	0	0	0	1	1	$105.00	$0.00	0.0
			TOTALS:			**6**	**5**	**5**	**6**	**1**	**0**	**1**	**24**	**$5,135.00**	**$0.00**	**0.0**

Printed on 01/27/2020 at 04:09 PM | * Stats based on Primary Demo **Page 1 of 2**

(a)

Contract # 26790171 Changes as of: 1/27/2020 at 11:42 AM Version: Original Order
CPE: 23/26/383 Flight: 2/17/20 - 2/23/20 Station: KTAB Con Type: POLITICAL/VOTE
Agency: DAVIS LENZ MEDIA Advertiser: Jon Francis Campaign Market: Abilene Total $: $5,135.00
6060 N CENTRAL EXPRESSWAY Product: POLITICAL Office: DALLAS Total Spots: 24
KATZ TELEVISION SUITE 560 Agency Order #: 9307478 Service: Nielsen Total CPP: $0.00
GROUP DALLAS, TX 75206 Buyer: Beth Davis, Haley Primary Demo: Total GRP:
Salesperson: ANDREA KRAUS Assistant: JACOB BURNETT Traffic #: 2357708
125 West 55th St New York, NY 10019 214-525-2617 214-525-2681
Comments: Separation: 15 Separation: Sep:15
2/17 - 2/23
Day/Time DP Program Rate Len 2/17 2/18 2/19 2/20 2/21 2/22 2/23 Total Spots Total $ $ CPP* GRP*
1 M-F 5:30a-7a KTAB Daybreak $145.00 30 1 1 1 1 1 0 0 5 $725.00 $0.00 0.0
2 M-F 7a-9a CBS This Morning $110.00 30 1 0 0 1 0 0 0 2 $220.00 $0.00 0.0
3 M-F 10a-11a The Price is Right $280.00 30 1 0 1 0 0 0 0 2 $560.00 $0.00 0.0
4 M-F 12n-12:30p KTAB Noon News $115.00 30 0 1 0 1 0 0 0 2 $230.00 $0.00 0.0
5 M-F 2p-3p Lets Make A Deal $105.00 30 1 0 1 0 0 0 0 2 $210.00 $0.00 0.0
6 M-F 4:30p-5p Jeopardy $115.00 30 0 0 1 0 0 0 0 1 $115.00 $0.00 0.0
7 M-F 5p-5:30p KTAB 5P News $190.00 30 1 1 0 1 0 0 0 3 $570.00 $0.00 0.0
8 M-F 6p-6:30p KTAB 6 P News $375.00 30 1 1 1 1 0 0 0 4 $1,500.00 $0.00 0.0
9 M-F 6:30p-7p Wheel of Fortune $450.00 30 0 1 0 1 0 0 0 2 $900.00 $0.00 0.0
10 Su 9:30a-10a Face the Nation $105.00 30 0 0 0 0 0 0 1 1 $105.00 $0.00 0.0
TOTALS: 6 5 5 6 1 0 1 24 $5,135.00 $0.00 0.0
Printed on 01/27/2020 at 04:09 PM | * Stats based on Primary Demo Page 1 of 2

(b)

Fig. 6.4: Comparison between the original document (a) and its corresponding HTML representation (b). Sample from VRDU benchmark.

This prompt sets the format and expectations for the model's response.

Following this, the instruction prompt provides specific directions for the task, guiding the language model to accurately extract and organize the necessary information:

> *"Given the following HTML table, extract key details and organize them into a single JSON object. Please provide values for the fields including 'advertiser', 'property', 'agency', 'tv_address', 'contract_num', 'product', 'gross_amount', 'flight_from', 'flight_to', and 'line_item' is an array (with 'channel', 'program_desc', 'program_end_date', 'program_start_date', and 'sub_amount'). Ensure that the extracted information accurately reflects the content of the HTML. Output must be JSON."*

With these instructions and the aforementioned HTML representation of the document, the final prompt is formed.

6.3.4 Implementation Details

We followed the best practices outlined in [152] and applied LoRA to all linear layers, including the query, key, and value layers within each attention block. In our implementation, we chose an alpha value of 64 to control the scale of low-rank updates applied to model weights. The rank of the low-rank matrices was set to 16, meaning the size of the trainable matrices. To reduce overfitting, we used a 0.05 dropout rate in the LoRA layers.

An NVIDIA A100 GPU was used for training with over 1,000 iterations. We used a baseline learning rate of 1×10^{-4} with a cosine learning rate scheduler. This scheduler gradually reduces the learning rate over the training period using a cosine curve, resulting in a smooth and controlled optimization process.

As shown in Figure 6.5, during inference, documents with multiple pages are split and individually converted to HTML duplicates, which are then fed into the LLM with the appropriate instruction prompt. Afterwards, we combine the predicted outcomes from all pages into a single JSON object with the structure that the evaluation code requires. To ensure proper evaluation, we sort the key values in both the JSON object and the ground truth.

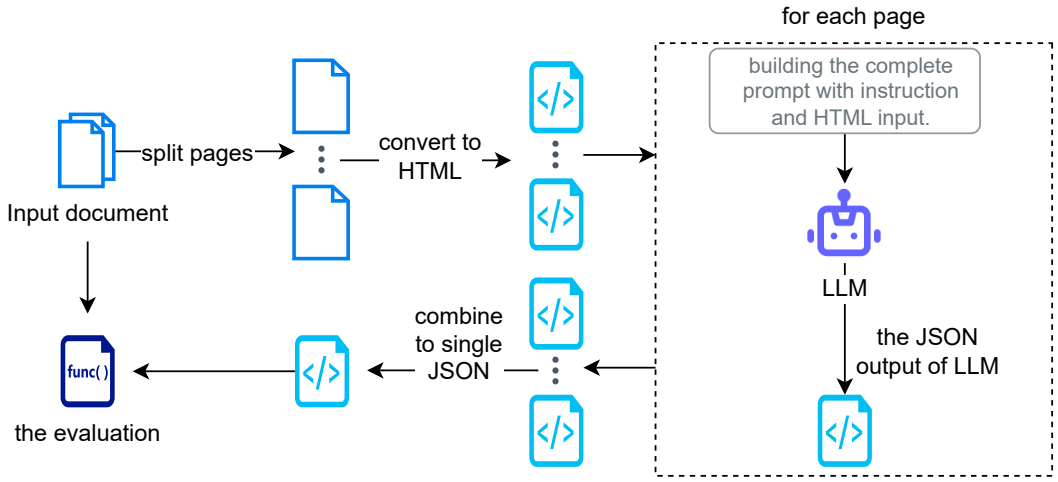


Fig. 6.5: Evaluation workflow: Documents are split into pages, converted to HTML, processed by the LLM with custom prompt, and outputs are compiled into a single JSON. The final output is compared with the ground truth for assessment.

6.4 Experiments and Results

In this section, we review the experiments and their results. First, we evaluate the Ad-buy dataset from VRDU under various settings defined by the benchmark. Next, we focus on a specific subset of the Ad-buy dataset, comprising only 100 training samples and a test set with unseen templates. This subset serves as challenging testing ground to conduct additional experiments. Using this subset, we assess the model's performance using a different input encoding, evaluate how the LLM performs without any training, and compare the results with another LLM. Lastly, we test the Consolidated Receipt Dataset (CORD) [153] to determine the versatility of our encoding approach.

6.4.1 Datasets

The VRDU benchmark includes two datasets: Ad-buy Forms and Registration Forms. The Registration Forms dataset has simpler layout with fewer details to extract. In contrast, the Ad-buy Forms dataset is more challenging, containing 641 documents primarily made up of invoices and receipts related to political advertisements. These documents have complex layouts with tables and detailed elements such as product names, flight dates, and total prices typical of invoices. The main challenge of this dataset is the accurate extraction and interpretation of structured data from these documents. This involves a variety of data types, including prices, dates, addresses,

and nested entities, as well as complex features like tables, multi-column layouts, and key-value pairs.

The benchmark provides high-quality OCR extraction results for text and their corresponding positions within the documents. It includes two tasks: Mixed Template Learning (MTL) and Unseen Template Learning (UTL). MTL evaluate the models' ability to handle various templates by incorporating multiple templates across training and testing sets. UTL evaluates the models' capacity to adapt to templates not seen during training. Each task in the VRDU dataset consists of 300 documents in the testing set, with four different training sets of 10, 50, 100, and 200 samples, respectively. This structure allows for assessing models on their efficiency with data and their performance with limited training data. Additionally, the authors implement a type-aware matching algorithm to accurately assess performance. The algorithm uses specific matching functions tailored to each entity's data type. For example, it employs numeric comparisons for monetary values to ensure that differences in formatting do not affect the matching results.

Additionally, the CORD [153] contains a thousand of Indonesian receipt images receipts. It comes with rich annotations for OCR and multi-level semantic labels for each word. The dataset is divided into training (800 receipts), validation (100 receipts), and test sets (100 receipts).

6.4.2 Evaluation on VRDU benchmark

Table 6.1 compares our proposed model with others, including LMDX, FormNet, and different versions of LayoutLM, evaluated on the Ad-buy dataset. It shows the performance of these models with varying data sizes and whether the templates in training/testing were mixed or unseen. The performance metrics include Micro-F1 and Line-Item F1 scores as defined in [154].

Our model shows significant improvement over the baseline methods such as the FormNet and the LayoutLM family in all settings. As the size of the training set increases, there is a consistent improvement in performance. The extraction of line items, which contain nested or itemized information, is particularly challenging because the evaluation process is strict; missing even a single item in a group results in it being marked completely incorrect.

Although the proposed method performs well, it has not reached the top performance achieved by LMDX due to several factors. First, LMDX has a larger architecture with greater processing capabilities. Additionally, LMDX benefits from pre-training on a

Tab. 6.1: Performance on Ad-buy dataset across various train sizes and template setting in train/test (mixed, unseen). The reported numbers are sourced from [143].

Size	Model	Mixed Template		Unseen
		Micro-F1	Line Item F1	Micro-F1
10	FormNet [141]	20.47	5.72	20.28
	LayoutLM [139]	20.20	6.95	19.92
	LayoutLMv2 [132]	25.36	9.96	25.17
	LayoutLMv3 [135]	10.16	5.92	10.01
	LMDX $_{\text{PaLM 2-S}}$ [143]	54.35	39.35	54.82
	Proposed	38.06	19.66	37.76
50	FormNet [141]	40.68	19.06	39.52
	LayoutLM [139]	39.76	19.50	38.42
	LayoutLMv2 [132]	42.23	20.98	41.59
	LayoutLMv3 [135]	39.49	19.53	38.43
	LMDX $_{\text{PaLM 2-S}}$ [143]	75.08	65.42	75.70
	Proposed	58.16	42.72	56.87
100	FormNet [141]	40.38	18.80	39.88
	LayoutLM [139]	42.38	21.26	41.46
	LayoutLMv2 [132]	44.97	23.52	44.35
	LayoutLMv3 [135]	42.63	22.08	41.54
	LMDX $_{\text{PaLM 2-S}}$ [143]	78.05	69.77	75.99
	Proposed	65.9	52.51	63.71
200	FormNet [141]	43.23	21.86	42.87
	LayoutLM [139]	44.66	23.90	44.18
	LayoutLMv2 [132]	46.54	25.46	46.31
	LayoutLMv3 [135]	45.16	24.51	44.43
	LMDX $_{\text{PaLM 2-S}}$ [143]	79.82	72.09	78.42
	Proposed	74.74	64.24	71.82

private dataset, enhancing its performance. LMDX also utilizes multiple inferencing techniques, leading to higher accuracy at a higher computational cost. Lastly, LMDX undergoes more training with 4,000 iterations compared to our 1,000 iterations. These factors, considering the computational cost and limitations in our experiments, explain the superior performance of the LMDX model in this context.

Table 6.2 shows the detailed performance of our model on the Ad-buy dataset for different fields, under different template sample frequencies (10, 50, 100, 200). For both *mixed* and *unseen* templates, as the number of samples increases, there is an improvement in F1 scores across most fields. Some fields such as 'gross amount', 'product', 'agency', and 'advertiser' consistently show higher F1 scores across both template types and all data sizes, indicating the model's effectiveness in these areas. Conversely, fields like 'tv address', 'line item', have lower F1 scores, especially in template with fewer samples, which means the model struggles more with these

Tab. 6.2: F1-Scores per field on the Ad-Buy dataset across various train sizes and template setting in train/test (mixed, unseen).

Template	Size	Advertiser	Agency	Contract Num	Flight From	Flight To	Gross Amount	Product	TV Address	Property	Line Item	Macro	Unrepeated entities	Micro
Mixed	10	82.19	76.18	76.75	71.09	72.55	88.57	84.64	44.77	67.85	19.66	68.43	74.21	38.06
Mixed	50	94.32	85.01	93.18	88.81	86.76	94.92	90.15	75.96	82.84	42.72	83.47	88.46	58.16
Mixed	100	94.14	88.05	95.17	91.41	92.08	96.63	93.2	79.17	86.98	52.51	86.94	91.15	65.9
Mixed	200	97.58	93.16	96.69	94.43	94.66	97.21	95.13	84.77	93.18	64.24	91.11	94.32	74.74
Unseen	10	78.48	71.95	77.64	73.48	73.55	90.85	83	42.77	68.33	19.33	67.94	73.9	37.76
Unseen	50	93.86	87.53	93.73	88.53	87.19	94.15	92.37	76.6	83.69	40.66	83.83	88.94	56.87
Unseen	100	94.24	91.81	93.78	89.29	90.42	96.02	93.32	78.99	86.58	49.26	86.37	90.63	63.71
Unseen	200	96.21	95.16	96.56	90.49	91.15	95.94	94.62	85.53	93.44	60.14	89.92	93.32	71.82

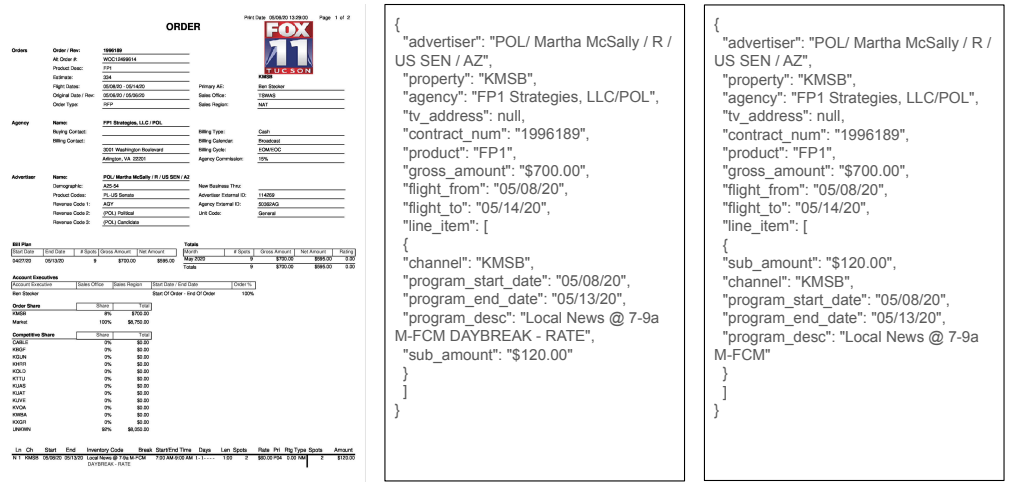


Fig. 6.6: (a) Sample image form VRDU. (b) Expected result. (c) Predicted result.

fields. The address field often contain multi-line text, while line item refers to groups of information presented in tabular form within these documents.

Figure 6.6 shows a sample document, ground truth, and model predictions. While most details are accurately extracted, there are instances where parts of the program description are missed. Such mistakes lead to a decline in the performance of the line-item.

6.4.3 Evaluation with Coordinate Embedding

As presented in [143], one encoding approach is to directly embed the normalized x|y coordinate pair of each word into the text input. As the authors state, this spatial

Tab. 6.3: Evaluation coordinate in text on unseen template 100 subset (F1-Scores).

Model	Advertiser	Agency	Contract Num	Flight From	Flight To	Gross Amount	Product	TV Address	Property	Line Item	Macro	Unrepeated entities	Micro
Coordinate	91.05	85.65	**96.04**	87.95	**90.54**	**96.51**	89.53	75	85.28	45.4	84.29	89.05	60.52
Proposed	**94.24**	**91.81**	93.78	**89.29**	90.42	96.02	**93.32**	**78.99**	**86.58**	**49.26**	**86.37**	**90.63**	**63.71**

Tab. 6.4: Evaluation zero-shot on unseen template 100 subset (F1-Scores).

Model	Advertiser	Agency	Contract Num	Flight From	Flight To	Gross Amount	Product	TV Address	Property	Line Item	Macro	Unrepeated entities	Micro
CodeLlama [146]	72.5	44.03	46.61	49.2	57.75	50.39	84.32	13.23	42.81	2.48	46.33	51.66	21.99
Proposed	**94.24**	**91.81**	**93.78**	**89.29**	**90.42**	**96.02**	**93.32**	**78.99**	**86.58**	**49.26**	**86.37**	**90.63**	**63.71**

context helps language models infer document layout relationships. For comparison, we train the LLM with this “coordinate-in-text” representation on a subset of dataset. As table 6.3 shows, our model generally outperforms the "coordinate" model in most fields, as indicated by higher F1 scores.

6.4.4 Zero-Shot Evaluation

In this experiment, we pass the input prompt to our LLM to evaluate its performance without tuning. Table 6.4 compares the proposed fine-tuned model against this zero-shot baseline on a subset of the VRDU dataset with unseen templates. We observe that certain information, such as advertiser and product names, can be extracted even without fine-tuning, due to their simple layout and straightforward availability for the LLM to detect. However, fine-tuning provides substantial gains, more than doubling scores across all categories by tailoring the model to the specific domain.

6.4.5 Evaluation of DeciLM-7B

To showcase the effectiveness of our encoding approach utilizing another LLM, we conducted a comparison with DeciLM-7B [155], a recently introduced instruction-following LLM that can handle long input contexts up to 8k. To ensure a fair comparison, we fine-tuned DeciLM-7B on the Ad-buy dataset using the same steps

Tab. 6.5: Evaluation DeciLM-7B on unseen template 100 subset (F1-Scores).

Model	Advertiser	Agency	Contract Num	Flight From	Flight To	Gross Amount	Product	TV Address	Property	Line Item	Macro	Unrepeated entities	Micro
DeciLM-7B [155]	90.79	86.13	**95.61**	85.56	**90.8**	**96.04**	90.35	78.3	82.83	46.35	84.28	88.85	61.11
Proposed	**94.24**	**91.81**	93.78	**89.29**	90.42	96.02	**93.32**	**78.99**	**86.58**	**49.26**	**86.37**	**90.63**	**63.71**

Tab. 6.6: Evaluation on CORD Dataset.

Size	Model	n-TED accuracy
50	Donut [142]	75.44
	LayoutLMv3LARGE [135]	87.29
	LMDX $_{\text{PaLM 2-S}}$ [143]	93.80
	proposed	89.9
800	Donut [142]	90.23
	LayoutLMv3LARGE [135]	96.21
	LMDX $_{\text{PaLM 2-S}}$ [143]	96.3
	proposed	91.4

as our proposed model. Table 6.5 presents the results, showing that both models perform similarly. However, in our specific application, CodeLlama generally outperforms DeciLM-7B.

6.4.6 Evaluation on CORD Dataset

We expanded our evaluation to include the CORD receipt dataset in two different settings: using only the first 50 samples to assess the model's few-shot learning capabilities, and using the complete dataset of 800 samples, in line with [143].

We followed the same procedure for prompt creation, training, and testing as applied to the VRDU dataset. However, this dataset presents an additional challenge as its lines in the OCR results are not aligned due to the presence of rotated and folded papers, making it difficult to construct the equivalent HTML.

Table 6.6 compares the n-TED accuracy [142] of various models on the CORD dataset, as reported in [143]. The results indicate that our model performs competitively in both training scenarios. With 50 samples, it achieves a higher n-TED accuracy compared to Donut and LayoutLMv3LARGE, but lower than LMDXPaLM 2-S. With 800 samples, the model's accuracy increases and remains higher than Donut.

6.5 Discussion

As we continue our research on using LLMs for various tasks, we encourage the community to adapt their inputs to be more easily understood by machines. LLMs that are familiar with HTML and JSON formats could be particularly advantageous. Our goal has been to demonstrate the practical utility of using HTML versions created from OCR data for information extraction with LLMs. For further accuracy, we recommend using any available proprietary software equipped with advanced features such as layout analysis, table detection, and superior OCR capabilities. Having additional metadata can greatly enhance the creation of precise HTML representation of a document.

The evaluation metric proposed by the dataset has a low tolerance for incomplete answers, as it does not accept partially correct responses. This might lead to an unfair comparison when a model's answer is marked incorrect due to missing a few characters.

6.6 Conclusion

This chapter explored the effective utilization of LLMs for structural information extraction from visually rich documents. Despite their impressive capabilities in natural language processing, LLMs often struggle with complex layouts, limiting their ability to accurately extract information. To address this challenge, a novel method was introduced, focusing on preparing input data to preserve the document's layout information.

The proposed approach involves transforming OCR outputs into structured HTML representations, which capture the spatial relationships and layout context of the document. This enables LLMs to understand not only the textual content but also its arrangement within the document, critical for accurate information extraction. The method leverages instruction-based prompting, where specific instructions and HTML representations are used as input for the LLM, fine-tuned to generate structured JSON outputs containing the extracted information.

The study utilized CodeLlama, a variant of Llama 2, designed to handle long contexts, understand HTML and JSON, and follow detailed instructions. The model was fine-tuned using PEFT with Low Rank Adaptation, reducing computational costs. Experiments on the VRDU benchmark demonstrated a significant improvement in

the LLM's ability to understand complex visual layouts and extract information accurately, with a more than 20 percent increase over baseline performances.

The findings highlight the effectiveness of the HTML representation and instruction-based prompting in improving LLM performance. The results also demonstrate the importance of input formatting and the choice of LLM in document understanding, with performance comparable to other SOTA methods. This research contributes to bridging the gap between the capabilities of LLMs and the practical requirements of visually rich document understanding, providing an efficient method for using LLMs in real world applications.

In summary, this research introduced a new approach for leveraging LLMs to extract information from visually rich documents with complex layouts. By converting OCR outputs into HTML format, the method preserves spatial layout and textual content, enabling LLMs to accurately extract information into a structured JSON format. The findings emphasize the importance of input formatting and the choice of LLM in document understanding, providing a valuable contribution to the field.

7 Zero-Shot Document Information Extraction using MLLM

> *The journey is the reward. And all of the trauma and mental health issues you accumulate along the way are the bonus prizes.*
>
> — **Dave Tarnowski**

7.1 Introduction

Multimodal Large Language Models (MLLMs) are a significant advancement in artificial intelligence, going beyond typical language models to integrate and understand a wide range of input types, including text, images, audio, and video. MLLMs process multiple modalities simultaneously to achieve a more comprehensive understanding that reflects real-world interactions, in contrast to unimodal models that are restricted to a single input type [156]. MLLMs are essentially LLM based models that can receive, reason about, and output multimodal information. Their development was inspired by the success of large language models, which relied on principles such as large scale pre-training, human like text generation, and few-shot learning capabilities. This has resulted in unexpected emergent capabilities in MLLMs, such as image captioning, visual question answering, and text recognition [157].

This chapter focuses on using powerful MLLMs to improve document understanding, particularly for visually rich documents. Understanding VRDs requires models to fully understand not only the words on the page but also their spatial layout and visual cues, such as tables, figures, and varying font sizes, in order to accurately extract and interpret information. We will look at how to extract structured information from various types of documents, effectively categorize their content, and understand the relationships between different document elements. Our discussion will center on the shortcomings of MLLMs and how they impact accuracy. We'll also include a

benchmark to assess how well models understand documents, as well as discuss the art of prompt engineering for document analysis.

7.2 Background

Traditional approaches to document information extraction have been limited by several kinds of restrictions. Rule based systems, while interpretable, necessitate extensive manual engineering and frequently fail to generalize across different document types. Template based extraction methods rely on known document structures, which limits their applicability to new formats. Supervised learning approaches, such as deep learning models, have shown significant improvements in extraction accuracy, but they require large annotated datasets for each document type. This results in a bottleneck in real-world deployment scenarios where new document types frequently emerge, and obtaining labeled data is costly and time consuming [157].

In contrast, LLM based key information extraction pipelines, particularly those utilizing prompt engineering, significantly reduce the system's reliance on large training datasets. They can quickly process industrial data with high accuracy without needing a lot of varied training examples, making them a practical approach for LLM and KIE applications in industry [158]. Furthermore, LLM based methods demonstrate superiority in handling OCR noise compared to traditional named entity recognition models. Integrating robust OCR tools with LLMs allows for more reliable data acquisition and can mitigate the impact of noise.

Despite their capabilities, using vision language models for document extraction faces several challenges. LLMs require sophisticated methods to ensure consistent quality and reliability in document information extraction tasks. Without proper prompting, these models frequently produce outputs with varying confidence levels and accuracy rates, making them unreliable for practical applications [156]. Maintaining consistency and coherence across extracted data is a continuous challenge for zero-shot document extraction systems. Developing robust consistency checking and error correction mechanisms is still a critical challenge for practical deployment of such systems.

Prompt engineering is the process of developing effective instructions, or "prompts", to guide a large language model. Prompting allows users to specify the processing and output formats for information extraction. This involves defining the required fields, their data types, and the overall structure, which helps the LLM understand the

task and the expected outcome. Essentially, the prompt acts as a blueprint, specifying how the model should interpret the input and then respond in a predefined format, like a JSON schema.

Zero-shot prompting refers to a model's ability to perform tasks without receiving any examples of those tasks during training. This capability is a valuable feature of LLMs and MLLMs, allowing them to generalize to new tasks with little or no task specific data. Instruction tuning further enhances zero-shot performance by framing tasks with clear instructions [157]. For instance, a zero-shot prompt for key information extraction might instruct the model to identify specific fields and their values from a document based purely on a textual description of what's needed.

While prompting reduces reliance on data, its zero-shot nature can result in a high number of errors, especially when the model is unfamiliar with the required output content types or conversions. Overall, research shows that providing detailed instructions improves performance and increases reliability. [159, 158].

Multimodal models integrating visual and textual inputs have demonstrated potential in image understanding [160], yet they frequently encounter difficulties when analyzing complex, text-rich document images. These models must concurrently identify text, interpret visual layout signals, and comprehend the spatial relationships among different document elements. Many MLLMs encounter difficulties in accurately capturing fine grained textual details present in images, as evidenced by benchmarks like TextVQA and OCR-VQA [156]. These benchmarks show how important it is for vision language architectures to have strong OCR capabilities.

Maintaining consistency across modalities is challenging, especially when there are differences between OCR text and image content [157]. The difficulty increases when documents, such as invoices, forms, tables, and technical diagrams, rely heavily on spatial arrangement to convey meaning. Recent research indicates that dividing visually complex documents into semantically coherent sections significantly improves extraction performance [161]. The BLOCKIE addresses these challenges by breaking down a document into "semantic blocks" and analyzing each one individually. A semantic block is a localized visual region containing text that can be interpreted independently of the rest of the document. This improves focused reasoning and improves generalization to novel layouts, resulting in more effective extraction of important information from complex documents.

Many of the current MLLMs analyze complete document images uniformly, presuming that all visual regions possess equal significance. This "one pass" approach

often impairs the focus on critical areas, resulting in inaccurate responses and hallucinations as the model attempts to navigate its inherent biases. To address this issue, the method proposed in [162] employs a coarse-to-fine reasoning manner. For datasets such as SROIE, relevant regions, or boxes, are initially identified based on the user's query. A blur reverse mask is then applied to enhance the focus of the MLLM on critical boxes while blurring the remaining parts of the image. The integration of targeted region selection and masking enables the MLLM algorithm to extract essential fields with greater precision from masked images, which decreases noise from unrelated areas and focuses computational resources on the document's most informative sections.

Document information extraction has moved from rule-based and template-based methods to more integrated systems. These systems use block segmentation, sophisticated prompt engineering, and text, image, and LLM driven reasoning. Each innovation improves previous methods, but cross modal consistency, spatial understanding, and error correction require further research and practical improvement.

7.3 Methodology

7.3.1 Baseline Prompt

To provide a clear comparison for our 3-phase framework, we use a unified prompt template. Table 7.1 contains the baseline prompt. This template is used with three different input modalities: text only, image only, and text plus image. The model receives the same set of instructions and output requirements in each case, but the format of the document varies depending on the baseline being evaluated—whether it is presented exclusively as OCR extracted text, exclusively as an image, or a combination of both.

The model is told that it is an expert in document analysis, and that its goal is to quickly identify the document's main topic, followed by extracting all factual and structured information into a valid JSON object. The JSON guidelines specify descriptive and self explanatory keys, tabular data under a dedicated "tables" key (with each table formatted as a nested object or array of objects), and the preservation of original formatting, including dates, currency symbols, and domain specific conventions, exactly as they appear in the source. The instructions also explicitly state that the model should not summarize the document, infer any data that is not explicitly present, and use a null value whenever a specific data point is missing or ambiguous.

Instruction: You are a highly skilled document analysis expert. Your objective is to thoroughly analyze a given document and perform two primary tasks:

- **Topic Identification:** Concisely determine the document's main topic.
- **Structured Information Extraction:** Extract all factual and structured information from the document, presenting it in a well formatted JSON object.

Follow these guidelines for the JSON output:

- Use descriptive and self explanatory keys for all data points (e.g., `"invoice_number"`, `"customer_name"`, `"total_amount"`).
- Represent tabular data under a `"tables"` key, with each table structured as either a nested object or an array of objects.
- Preserve the original formatting of values, including dates, currency symbols, and other document specific formats.
- If specific data points are absent or ambiguous, include the corresponding key with a value of `null`.
- Ensure that the JSON output is valid and well structured, with no syntax errors or missing commas.

Important Considerations:

- Do not summarize the document.
- Do not fabricate or infer information.
- Extract only data that is explicitly present in the document.

Tab. 7.1: Baseline prompt for expert level analysis of insurance and legal documents. This structured instruction is tailored for multimodal models handling document understanding tasks.

Finally, the prompt emphasizes that the output JSON must be syntactically correct, with no missing commas or mismatched braces.

7.3.2 3-Phase Zero-Shot Extraction Framework

To address the aforementioned issues, we propose a 3-phase approach that divides the complex task of document information extraction into manageable, sequential steps. This framework builds on the strengths of modern vision language models

while incorporating quality control mechanisms. The pipeline has three distinct phases:

1. **Document Analysis and Block Identification**: This phase focuses on high level document comprehension and structural analysis, identifying the dominant language and segmenting the document into coherent units of related information.

2. **Schema Driven Extraction**: The dynamic JSON schemas generated in Phase 1 guide the extraction process in Phase 2, resulting in well structured data.

3. **Consolidation and Verification**: The verification mechanism in Phase 3 increases robustness by catching errors that may have slipped through the initial pass, ensuring the accuracy and reliability of the extracted data.

This zero-shot framework offers flexibility no need for per document template design and reliability scheme enforcement and self review, making it a dependable solution for document processing systems.

Phase 1: Document Analysis and Block Identification

The first phase of our framework focuses on obtaining a broad understanding of the document's structure and content. Rather than attempting to extract data right away, this phase lays the groundwork for future operations with several key steps.

Initially, the model places the document in a broad category, such as invoice, contract, or resume. This categorization allows downstream steps to use content and formatting expectations that are specific to the category. The model then divides the document's content into discrete "information blocks", each representing a coherent unit of related information (such as "sender address", "invoice details", or "product table"). As shown in Table 7.2, each block is assigned a descriptive label, a brief textual description, and a structural hint indicating the expected format (key value pairs, table, list, free form text, or another category). The expected JSON schema is provided as input, along with the prompt.

Phase 2: Data Extraction

Phase 2 automatically creates block specific JSON schemas based on assigned type hints after Phase 1 produces labeled blocks and concludes the document is ready for further processing. By defining each block's expected structure in JSON, the

system ensures that all subsequent outputs follow a predictable format suitable for downstream consumption. During extraction, the model is given carefully crafted prompts, as detailed in Table 7.3, that provide specific instructions on how to handle each data type. The schema specifies field extraction for key value blocks, row column structures for tables, and output format for narrative or list style content (e.g., single string or array). The process preserves original formatting, such as dates, currency symbols, and domain-specific conventions, to ensure that the values in the resulting JSON match those in the document. By strictly adhering to the dynamically generated schema, the model avoids common errors like missing required fields or producing malformed JSON.

Instruction: You are a document analysis expert. Analyze the provided document (image and OCR text). Your tasks are:
1. **Document Language Identification**: Identify the language of the OCR TEXT.
2. **Topic Identification (Summary and Classification)**:

- Provide a concise 1-2 sentence `summary` of the document's main topic or purpose.
- Provide a `classification` (document type like Invoice, Contract, Resume).

3. **Information Block Segmentation**: Segment the *entire content* into distinct `information_blocks`. For each block:

- `block_name`: English camelCase name (e.g., "invoiceDetails", "senderAddress").
- `description`: Brief description of the block's content.
- `data_type_hint`: (e.g., "key_value_pairs", "table").
- Include an `otherRelevantInfo` block for significant remaining text, with description and hint `free_form_text_summary`.

4. **Output Format**: Return a SINGLE, VALID JSON object. The top-level keys must be `dominant_language_detected`, `summary`, `classification`, `is_relevant`, and `information_blocks`. Adhere strictly to the provided JSON schema.

Tab. 7.2: Prompt for Phase 1: Document Understanding and Block Identification. This structured instruction is tailored for multimodal models handling document understanding tasks.

Phase 3: Consolidation and Verification

Phase 3 introduces an additional LLM based verification step. This self correction mechanism asks the language model to compare its own initial extraction results against the original document. As outlined in Table 7.4, it actively corrects any inaccuracies and fills in missing information, all while strictly adhering to the established block structure. The verification process has three goals. It corrects extraction errors by comparing with the original source, incorporates block-level description information, and preserves the document's formatting.

The results of the previous phases are organized into a cohesive final structure after full verification. This organization provides detailed extraction status tracking,

Instruction: You are an expert data extractor.
Document Context: Type=`{doc_classification}`, Summary=`{doc_summary}`, Detected Language=`{doc_language_detected}`.
Your task is to extract structured information from the provided document (image and OCR text) for several pre-identified blocks of information.
General Extraction Guidelines (these apply to the content extracted for EACH block): When extracting data for the current block:

- Use descriptive and self-explanatory keys for all data points (e.g., `invoiceNumber`, `customerName`, `totalAmount`).
- Preserve the original formatting of values as much as possible, including dates, currency symbols, and other document-specific formats.
- Ensure your JSON output for this block is valid and well-structured.
- Do not fabricate or infer information. Extract only data that is explicitly present in this block of the document.

Here are the blocks you need to process: `{blocks_to_process_str}`
Your entire response MUST be a single JSON object. The top-level keys of this JSON object MUST be the exact 'Block Name's listed above.
The value for each `block_name` key MUST conform to the expected structure for its 'Expected Content Type Hint' (e.g., an object of key-values for `"key_value_pairs"`, an array of objects for `"table"`, an array for `"list_of_items"`, or an object for `"single_value_text"` or `"free_form_text_summary"`).
Adhere strictly to the provided overall JSON schema which defines this structure.

Tab. 7.3: Prompt for Phase 2: Consolidated Data Extraction. This structured instruction is tailored for multimodal models handling document understanding tasks.

showing which blocks were processed and which failed. This transparency allows downstream systems to make informed decisions regarding data completeness and reliability.

Instruction: You are a meticulous Quality Assurance specialist. An initial automated extraction has been performed on a document.
The initial extraction attempt for various document blocks is provided below as "**Initial Extracted Data**".
Your task is to review this "Initial Extracted Data" against the "Original Document" (image and OCR text). Focus on:

1. **Accuracy**: Are the extracted values correct according to the document? Correct any errors.
2. **Completeness**: Is any relevant information missing from a block that should have been extracted based on its description and the document content? If so, add the missing key value pairs *within the existing block structure*.
3. **Formatting**: Ensure values preserve original formatting (dates, currencies, etc.) where appropriate.
4. **Structure**: Maintain the existing block structure. If a block was extracted as an object, keep it an object. If it was a list, keep it a list. Add or correct fields *within* these existing block structures. You may add new blocks if necessary, but do not remove any existing blocks.

Output: A **corrected and completed version** of the JSON data, maintaining the same overall block structure as the `"Initial Extracted Data"`.

- If the initial extraction for a block is perfect, include that block as-is in your output.
- If no changes are needed for the entire `"Initial Extracted Data"`, simply return it as is.

Your output should be a single JSON object representing the reviewed data, with the same top-level keys as the `"Initial Extracted Data"` (which are the block names).

Tab. 7.4: Prompt for Phase 3: Consolidation and Verification. This structured instruction is tailored for multimodal models handling document understanding tasks.

7.4 Experiments and Results

For evaluation, we selected two of the most recent SOTA MLLMs: InternVL3 8B and Gemma3 12B. These models were chosen to assess how different underlying MLLM capabilities impact the performance of our proposed zero-shot document information extraction framework.

InternVL3 [163] is an open-source MLLM capable of understanding and reasoning across text, images, and video. It represents a significant step forward for the InternVL family, incorporating architectural and training methodology innovations that enable superior performance across a wide range of multimodal tasks. InternVL3 uses a "ViT-MLP-LLM" architecture and pre-trained language models like Qwen2.5 to perform a variety of tasks, including document analysis, video reasoning, and multilingual comprehension. Figure 7.1 shows the performance of the InternVL3 family.

Gemma3 [164] is Google's newest generation of lightweight, open-weight AI models, designed for efficient deployment and diverse multimodal tasks. Building on Gemma2 [165], it integrates text and image inputs. It has an extended context window (up to 128k tokens) enabled by optimized local global attention and upgraded Rotary Positional Embeddings. Available in four parameter sizes (1B, 4B, 12B, 27B) with multilingual support, Gemma3 significantly outperforms its predecessors, with the 27B model matching Gemini 1.5-Pro in benchmarks [164]. This makes it ideal for applications like document analysis and visual question answering.

Instruction:
You are a document analysis expert tasked with evaluating the accuracy and completeness of four JSON files, each representing the content extracted from the same PDF document.
For each JSON file, identify: numbers of errors in the extracted valuse.
Explanation: do not evaluate based on structure but rather on the accuracy of the information it contains.
Provide your final response in JSON format.

Tab. 7.5: Prompt for Grand truth evaluation.

Tab. 7.6: Error counts per document using the InternVL3 for text-only, image-only, image+text, and proposed three-phase methods

Sample	Baseline text	Baseline image	Baseline image + text	Proposed
cv_1	3	30	4	5
cv_2	5	18	8	**4**
cv_3	4	14	0	1
cv_4	7	10	5	**3**
cv_5	13	50	7	**7**
cv_6	4	17	6	5
cv_7	3	12	5	6
cv_8	3	16	2	**1**
cv_9	7	18	5	9
cv_10	3	25	2	5
Insurance_1	1	8	2	10
Insurance_2	5	12	5	**5**
Insurance_3	7	12	4	**0**
Insurance_4	8	12	6	14
Insurance_5	4	10	3	**2**
Insurance_6	5	20	6	**1**
Insurance_7	10	15	11	**7**
Insurance_8	3	10	2	5
Insurance_9	5	16	4	6
Insurance_10	9	9	9	**9**
invoice_1	12	11	14	**10**
invoice_2	7	40	4	12
invoice_3	10	12	12	15
invoice_4	10	14	8	9
invoice_5	12	9	4	**3**
invoice_6	4	10	5	6
invoice_7	4	15	4	**4**
invoice_8	5	12	8	10
invoice_9	3	9	3	5
invoice_10	2	8	8	5
ticket_1	4	12	2	4
ticket_2	6	7	9	10
ticket_3	8	8	8	**8**
ticket_4	4	17	5	**3**
ticket_5	7	16	4	11
ticket_6	10	12	3	6
ticket_7	9	9	6	**2**
ticket_8	3	8	6	**0**
ticket_9	0	10	1	7
ticket_10	4	18	6	10
average	5.825	14.775	5.4	6.125

Tab. 7.7: Error counts per document using the Gemma3 for text-only, image-only, image+text, and proposed three-phase methods

Sample	Baseline text	Baseline image	Baseline image + text	Proposed
cv_1	1	8	5	2
cv_2	8	14	8	**4**
cv_3	7	16	4	**3**
cv_4	7	12	6	**5**
cv_5	4	8	4	**2**
cv_6	3	12	2	**1**
cv_7	2	15	2	**2**
cv_8	8	18	4	**4**
cv_9	4	16	3	**1**
cv_10	2	7	**1**	**1**
Insurance_1	9	15	5	**2**
Insurance_2	2	10	**1**	**1**
Insurance_3	6	7	6	**2**
Insurance_4	11	26	13	**10**
Insurance_5	4	18	2	4
Insurance_6	8	15	4	**4**
Insurance_7	3	16	7	6
Insurance_8	1	14	1	**0**
Insurance_9	4	18	7	**2**
Insurance_10	10	20	6	**5**
invoice_1	10	12	5	**2**
invoice_2	20	27	15	**10**
invoice_3	9	15	6	**4**
invoice_4	20	12	10	**6**
invoice_5	7	22	9	**2**
invoice_6	6	18	5	**4**
invoice_7	2	8	1	3
invoice_8	4	9	3	**3**
invoice_9	5	11	1	**1**
invoice_10	7	15	4	5
ticket_1	2	12	1	2
ticket_2	7	16	6	11
ticket_3	2	20	5	3
ticket_4	0	2	0	**0**
ticket_5	7	7	3	**2**
ticket_6	6	14	7	**5**
ticket_7	7	9	8	**4**
ticket_8	8	8	3	**2**
ticket_9	1	8	4	**1**
ticket_10	4	11	2	**0**
average	5.95	13.525	4.725	3.275

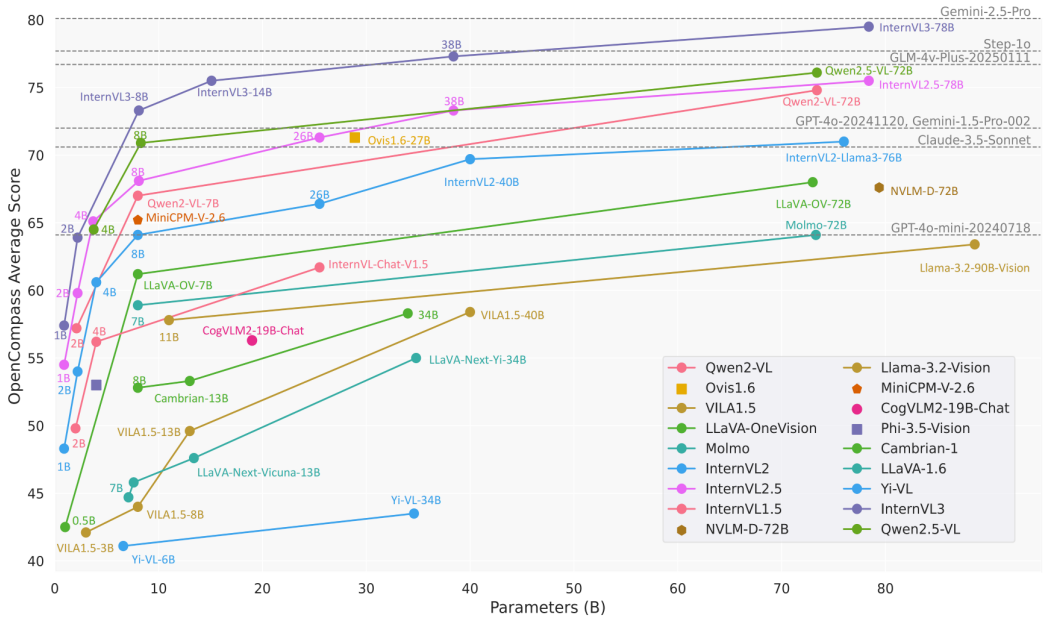


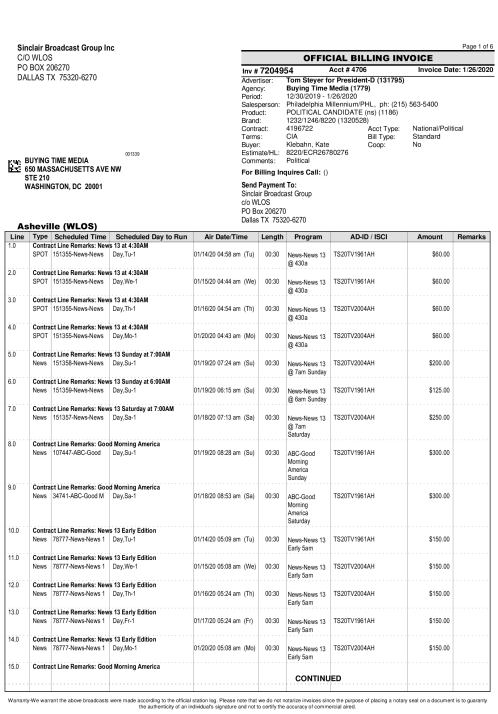
Fig. 7.1: Performance of various MLLMs on the OpenCompass multimodal academic leaderboard. Reproduced from [163].

7.4.1 Dataset

Current benchmarks tend to focus on a narrow range of tasks or datasets, which makes qualitative comparisons very hard[159]. Identifying appropriate metrics for evaluating zero-shot performance is an additional challenge. In response to this issue, we created a dataset of 40 real-world documents with various layouts and content, divided into four categories: resumes, insurance documents, invoices, and tickets. The documents are editable PDF format, allowing evaluation, and Tesseract [149] was used for OCR processing. This diverse dataset enables a comprehensive evaluation of the model's zero-shot content understanding and the performance gains achieved through targeted prompting strategies. The selected document types encompass a diverse array of layouts and information structures, including semi-structured data in invoices, unstructured text in resumes, and complex tabular formats in insurance forms, each illustrating what is required in practical enterprise applications. Figure 7.2 shows a few sample documents from our dataset that demonstrate the variety.

7.4.2 Results

Our evaluation method uses GPT-4o-mini, an independent LLM [166], to assess information extraction accuracy across various approaches. Reasoning models like



Sinclair Broadcast Group Inc
C/O WLOS
PO BOX 206270
DALLAS TX 75320-6270

Page 1 of 6

OFFICIAL BILLING INVOICE

Inv # 7204954 Acct # 4706 Invoice Date: 1/26/2020

Advertiser: Tom Steyer for President-D (131795)
Agency: Buying Time Media (1779)
Period: 12/30/2019 - 1/26/2020
Salesperson: Philadelphia Millennium/PHL, ph: (215) 563-5400
Product: POLITICAL CANDIDATE (ns) (1186)
Brand: 1232/1246/8220 (1320526)
Contract: 4196722 Acct Type: National/Political
Terms: CIA Bill Type: Standard
Buyer: Klebahn, Kate Coop: No
Estimate/HL: 8220/ECR/26780276
Comments: Political

For Billing Inquires Call: ()

Send Payment To:
Sinclair Broadcast Group
c/o WLOS
PO Box 206270
Dallas TX 75320-6270

BUYING TIME MEDIA
650 MASSACHUSETTS AVE NW
STE 210
WASHINGTON, DC 20001

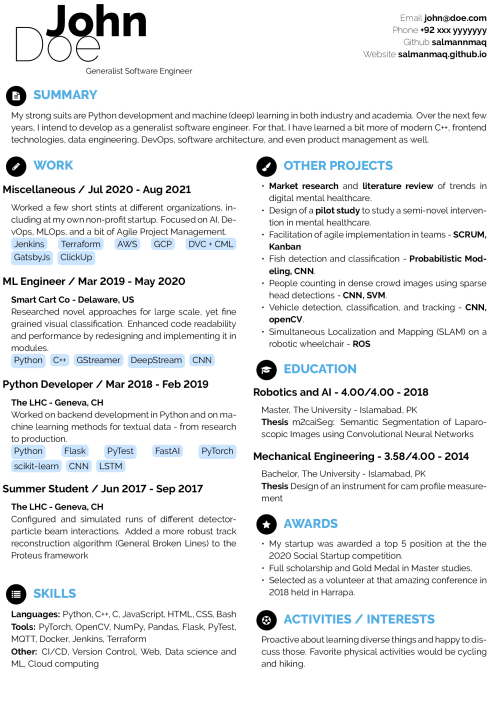
Asheville (WLOS)

Line	Type	Scheduled Time	Scheduled Day to Run	Air Date/Time	Length	Program	AD-ID / ISCI	Amount	Remarks
1.0		Contract Line Remarks: News 13 at 4:30AM							
	SPOT	151355-News-News	Day,Tu-1	01/14/20 04:58 am (Tu)	00:30	News-News 13 @ 430a	TS20TV1961AH	$60.00	
2.0		Contract Line Remarks: News 13 at 4:30AM							
	SPOT	151355-News-News	Day,We-1	01/15/20 04:44 am (We)	00:30	News-News 13 @ 430a	TS20TV1961AH	$60.00	
3.0		Contract Line Remarks: News 13 at 4:30AM							
	SPOT	151355-News-News	Day,Th-1	01/16/20 04:54 am (Th)	00:30	News-News 13 @ 430a	TS20TV2004AH	$60.00	
4.0		Contract Line Remarks: News 13 at 4:30AM							
	SPOT	151355-News-News	Day,Mo-1	01/20/20 04:43 am (Mo)	00:30	News-News 13 @ 430a	TS20TV2004AH	$60.00	
5.0		Contract Line Remarks: News 13 Sunday at 7:00AM							
	News	151358-News-News	Day,Su-1	01/19/20 07:24 am (Su)	00:30	News-News 13 @ 7am Sunday	TS20TV2004AH	$200.00	
6.0		Contract Line Remarks: News 13 Sunday at 6:00AM							
	News	151359-News-News	Day,Su-1	01/19/20 06:15 am (Su)	00:30	News-News 13 @ 6am Sunday	TS20TV1961AH	$125.00	
7.0		Contract Line Remarks: News 13 Saturday at 7:00AM							
	News	151357-News-News	Day,Sa-1	01/18/20 07:13 am (Sa)	00:30	News-News 13 @ 7am Saturday	TS20TV2004AH	$250.00	
8.0		Contract Line Remarks: Good Morning America							
	News	107447-ABC-Good	Day,Su-1	01/19/20 08:28 am (Su)	00:30	ABC-Good Morning America Sunday	TS20TV1961AH	$300.00	
9.0		Contract Line Remarks: Good Morning America							
	News	34741-ABC-Good M	Day,Sa-1	01/18/20 08:53 am (Sa)	00:30	ABC-Good Morning America Saturday	TS20TV1961AH	$300.00	
10.0		Contract Line Remarks: News 13 Early Edition							
	News	78777-News-News 1	Day,Tu-1	01/14/20 05:09 am (Tu)	00:30	News-News 13 Early 5am	TS20TV1961AH	$150.00	
11.0		Contract Line Remarks: News 13 Early Edition							
	News	78777-News-News 1	Day,We-1	01/15/20 05:08 am (We)	00:30	News-News 13 Early 5am	TS20TV2004AH	$150.00	
12.0		Contract Line Remarks: News 13 Early Edition							
	News	78777-News-News 1	Day,Th-1	01/16/20 05:24 am (Th)	00:30	News-News 13 Early 5am	TS20TV2004AH	$150.00	
13.0		Contract Line Remarks: News 13 Early Edition							
	News	78777-News-News 1	Day,Fr-1	01/17/20 05:24 am (Fr)	00:30	News-News 13 Early 5am	TS20TV1961AH	$150.00	
14.0		Contract Line Remarks: News 13 Early Edition							
	News	78777-News-News 1	Day,Mo-1	01/20/20 05:08 am (Mo)	00:30	News-News 13 Early 5am	TS20TV2004AH	$150.00	
15.0		Contract Line Remarks: Good Morning America							

CONTINUED

Warranty-We warrant the above broadcasts were made according to the official station log. Please note that we do not notarize invoices since the purpose of placing a notary seal on a document is to guaranty the authenticity of an individual's signature and not to certify the accuracy of commercial aired.

(a) Invoice



John Doe
Generalist Software Engineer

Email john@doe.com
Phone +92 xxx yyyyyyy
GitHub salmanmaq
Website salmanmaq.github.io

SUMMARY

My strong suits are Python development and machine (deep) learning in both industry and academia. Over the next few years, I intend to develop as a generalist software engineer. For that, I have learned a bit more of modern C++, frontend technologies, data engineering, DevOps, software architecture, and even product management as well.

WORK

Miscellaneous / Jul 2020 - Aug 2021

Worked a few short stints at different organizations, including at my own non-profit startup. Focused on AI, DevOps, MLOps, and a bit of Agile Project Management.
Jenkins Terraform AWS GCP DVC + CML GatsbyJs ClickUp

ML Engineer / Mar 2019 - May 2020

Smart Cart Co - Delaware, US
Researched novel approaches for large scale, yet fine grained visual classification. Enhanced code readability and performance by redesigning and implementing it in modules.
Python C++ GStreamer DeepStream CNN

Python Developer / Mar 2018 - Feb 2019

The LHC - Geneva, CH
Worked on backend development in Python and on machine learning methods for textual data - from research to production.
Python Flask PyTest FastAI PyTorch scikit-learn CNN LSTM

Summer Student / Jun 2017 - Sep 2017

The LHC - Geneva, CH
Configured and simulated runs of different detector-particle beam interactions. Added a more robust track reconstruction algorithm (General Broken Lines) to the Proteus framework

SKILLS

Languages: Python, C++, C, JavaScript, HTML, CSS, Bash
Tools: PyTorch, OpenCV, NumPy, Pandas, Flask, PyTest, MQTT, Docker, Jenkins, Terraform
Other: CI/CD, Version Control, Web, Data science and ML, Cloud computing

OTHER PROJECTS

- **Market research** and **literature review** of trends in digital mental healthcare.
- Design of a **pilot study** to study a semi-novel intervention in mental healthcare.
- Facilitation of agile implementation in teams - **SCRUM, Kanban**
- Fish detection and classification - **Probabilistic Modeling, CNN**.
- People counting in dense crowd images using sparse head detections - **CNN, SVM**.
- Vehicle detection, classification, and tracking - **CNN, openCV**.
- Simultaneous Localization and Mapping (SLAM) on a robotic wheelchair - **ROS**

EDUCATION

Robotics and AI - 4.00/4.00 - 2018

Master, The University - Islamabad, PK
Thesis m2calSeg: Semantic Segmentation of Laparoscopic Images using Convolutional Neural Networks

Mechanical Engineering - 3.58/4.00 - 2014

Bachelor, The University - Islamabad, PK
Thesis Design of an instrument for cam profile measurement

AWARDS

- My startup was awarded a top 5 position at the the 2020 Social Startup competition.
- Full scholarship and Gold Medal in Master studies.
- Selected as a volunteer at that amazing conference in 2018 held in Harrapa.

ACTIVITIES / INTERESTS

Proactive about learning diverse things and happy to discuss those. Favorite physical activities would be cycling and hiking.

(b) Resume

mawista

MAWISTA GmbH * Esslinger Str. 83 * 73207 Plochingen * Deutschland

Mohammad Minouei

Am Harzhübel 112
67663 Kaiserslautern
Deutschland

Es betreut Sie / *Your contact:*
MAWISTA GmbH
Versicherungsvermittlung
Esslinger Str. 83
73207 Plochingen
Deutschland
Tel.: +49 7024 469 51-0
Fax-Nr.: +49 7024 469 51-20
E-Mail: info@mawista.com
Internet: www.mawista.com

Ausstellungsdatum / *Date of issue:*
21.10.2024

Bescheinigung über privaten Krankenversicherungsschutz für die Erteilung von Aufenthaltstiteln / *Confirmation for authorities*

Tarif / *tariff* MAWISTA Student Classic Plus
Versicherungsschein-Nr. / *Policy no.* MAW76785577
Please refer to this number for all future correspondence

Der Versicherer Allianz Partners - AWP P&C S.A., Niederlassung für Deutschland, bescheinigt hiermit, dass für die nachstehend genannte Person Krankenversicherungsschutz besteht:

Versicherte Person / *Insured person*	Geburtsdatum / *Date of birth*	Pass-Nr. / *Passport no.*
Mohammad Minouei	31.01.1992	N97189852
Versicherungsbeginn / *Start of insurance*	**Versicherungsende** / *End of insurance*	
01.11.2024	31.10.2025	

Die Selbstbeteiligung je tarifischer Leistung bei ambulanter Behandlung und bei Zahnbehandlungen beträgt im Tarif Classic 20 %, im Tarif Classic Plus 15 % und im Tarif Comfort 10 %, Maximal beträgt die Selbstbeteiligung pro Kalenderjahr € 250,–.

Der Versicherungsschutz umfasst folgende Leistungen:
- Ambulante Behandlung, einschließlich Verband- und Arzneimittel
- Zahnärztliche Versorgung, schmerzstillende Zahnbehandlung
- Stationäre Heilbehandlung inklusive Rehabilitationsmaßnahmen als Anschlussheilbehandlung
- Behandlungen für psychische Erkrankungen

Im Falle einer Transportunfähigkeit besteht Versicherungsschutz auch über das Versicherungsende hinaus, maximal jedoch für 6 Wochen.

Der Versicherungsschutz erfüllt die Anforderungen gemäß der Verordnung (EG) Nr. 810/2009 des Europäischen Parlaments und des Rates vom 13.Juli 2009 und gilt in allen Mitgliedstaaten, die den Schengen Besitzstand in vollem Umfang anwenden. Die Versicherungssumme ist auf die Mindestdeckung von € 30.000,- nicht begrenzt.

Der genaue Versicherungsumfang ist den Versicherungsbedingungen zum Produkt MAWISTA Student zu entnehmen.

Mit freundlichen Grüßen / *Yours sincerely*
für den Versicherer / *on behalf of the insurer:*
AWP P.C.S.A.

Jacob Fuest
Hauptbevollmächtigter

AWP P&C S.A.
Niederlassung für Deutschland
Bahnhofstraße 16
D-85609 Aschheim
Hauptbevollmächtigter: Jacob Fuest

Registergericht: München HRB 4605
USt.-IdNr: DE 129274528
VersSt.-Nr.: 802/V90802001910
Commerzbank München
IBAN: DE39 7008 0000 0902 5406 06

AWP P&C S.A.
Aktiengesellschaft französischen Rechts
Sitz der Gesellschaft: Saint-Ouen (Frankreich)
Handelsregister: R.C.S. Bobigny 519 490 080
Vorstandsvorsitzende: Sirma Boshnakova

(c) Insurance

DB
CIV 1080

Online-Ticket

ICE Fahrkarte

Gültigkeit: **19.08.2024 00:00 Uhr bis 21.08.2024 03:00 Uhr**
Sie können alle Züge nutzen, die auf Ihrer Fahrkarte angegeben sind. Für Züge des Nahverkehrs z.B. RE, RB, S, besteht keine Zugbindung.
Gilt in der Schweiz nur in gebuchten Zügen gemäß Reservierung.

Super Sparpreis Europa (Einfache Fahrt)

Klasse	**2. Klasse**
Reisender	**1 Person (27-64 Jahre) (32 Jahre)**
Einfache Fahrt	**Kaiserslautern Hbf → Interlaken Ost**
	Via: <1080>NW°WOE°KA°RA°BAD°OG°FR°BAS
	<1185>Olten°Bern
Zugbindung	**BUS 12009, 09:20 Uhr am 19.08.2024**
	ICE 275, 10:12 Uhr am 19.08.2024
	ICE 275, 11:56 Uhr am 19.08.2024

Eine Stornierung Ihrer Fahrkarte ist ausgeschlossen.

Gesamtpreis 36,90 €. Gebucht am 31.05.2024 um 13:58 Uhr.
Dieses Dokument ist nicht vorsteuerabzugsfähig.

Barcode bitte nicht knicken!

5726 4 9038978

Zangenabdruck

Mohammad Minouei
Auftragsnummer: 572649038978

Ihre Reiseverbindung und Reservierung - Einfache Fahrt am 19.08.2024

Halt	Datum	Zeit	Gleis	Produkte	Reservierung / Hinweise
Kaiserslautern Hbf	19.08.	ab 07:32	5	S 1	
Neustadt(Weinstr)Hbf	19.08.	an 08:00	4		
Neustadt(Weinstr)Hbf	19.08.	ab 08:09	5	RE 6 (12015)	
Karlsruhe Hbf	19.08.	an 08:54			
Karlsruhe Hbf Südausgang	19.08.	ab 09:20		BUS12009	
Bahnhof, Baden-Baden	19.08.	an 09:55	2		
Baden-Baden	19.08.	ab 10:12		ICE 275	
Interlaken Ost	19.08.	an 13:58	5	ICE 275	

Wichtige Nutzungshinweise:
- Ihre Fahrkarte ist nur gültig mit einem amtlichen Lichtbildausweis. Dieser ist bei der Kontrolle vorzuzeigen.
- Bei Fahrkarten mit BahnCard-Rabatt zeigen Sie bitte zusätzlich Ihre gültige BahnCard vor.
- Es gelten die nationalen und internationalen Beförderungsbedingungen der DB AG. Innerhalb von Verkehrsverbünden und Tarifgemeinschaften gelten deren Bestimmungen. Alle Bedingungen finden Sie unter www.bahn.de/agb und www.diebefoerderer.de.
- Eine Fahrkarte entspricht grundsätzlich einem Beförderungsvertrag, mehrere Fahrkarten mehreren Beförderungsverträgen. Vertraglicher Beförderer können dabei ein oder mehrere Verkehrsunternehmen sein. Es handelt sich bei dieser Fahrkarte um eine Durchgangsfahrkarte gemäß Europäischer Fahrgastrechte-Verordnung für den Eisenbahnverkehr.
- Kleinkindabteile sind bei Bedarf für diese Personengruppe zu räumen.

Bitte informieren Sie sich kurz vor Reisebeginn auf unserer Website oder in der App, ob kurzfristige Fahrplanänderungen vorliegen. Wir danken Ihnen für Ihre Buchung und wünschen eine angenehme Reise.

Mohammad Minouei
19 08

DB
Dir ist wichtig, die Zukunft mitzugestalten?
Finde deinen Job im Team DB.
db.jobs

Ticketcode: LU1ABD6Q

Seite 1 / 1

(d) Ticket

Fig. 7.2: Sample documents from the collected dataset. Each image represents one of the four main categories used for evaluating zero-shot information extraction.

Tab. 7.8: Comparison of error counts and detailed issues across different extraction methods for sample `invoice_5`

Method	Errors Count	Issues
Baseline text	7	Line 5 & 6 amounts missing leading "$"; Line 7 amount mis-recorded as $1,250.00 (should be $250.00); Lines 8 & 9 program/AD-ID fields merged incorrectly; Line 10 AD-ID and amount merged/missing "$"; Table headers concatenated (Scheduled Time/Scheduled Day/Air Date/Length combined); Important "remarks" text sometimes lumped into wrong column.
Baseline image	22	Account number wrong ("4786" vs "4706"); Advertiser name & ID mangled; Agency name & ID wrong; Period and invoice date do not match source; Client/brand fields do not correspond to original; Buyer and terms fields invented ("CIA", "Net 30"); Estimate number wrong; Line items completely mismatched (dates, programs, AD-IDs, amounts); Total amount missing.
Baseline image + text	9	Column header typo ("...Length}"); Line 7 amount mis-recorded as $1,250.00 (should be $250.00); Program name truncated for Good Morning America; Remarks for lines 10–14 mis-transcribed ("Early Sam" vs "Early 5 am"); Some fields (e.g., "Scheduled Day to Run Air Date/Time Length}") mis-split.
Proposed	2	Agency street number wrong (555 vs 650 Massachusetts Ave NW); Line 7 amount mis-recorded as $1,250.00 (should be $250.00).

GPT-4o-mini are ideal for detecting subtle errors and comparing complex information structures. Table 7.5 describes that GPT-4o-mini functions as a document analysis expert tasked with comparing four separate JSON output files, each representing extractions from the same document, to the corresponding PDF document. The model's primary function is to detect and quantify errors in the values extracted from each JSON file. Importantly, the evaluation focuses solely on the accuracy of the informational content, ignoring structural differences in the JSON formatting. The model's final output is presented in a standard JSON format, allowing for more consistent and efficient analysis.

When utilizing InternVL3 as the underlying MLLM (Table 7.6), the image-only baseline performs the worst with a mean of 14.775 errors per document, confirming that visual features alone are insufficient for text rich documents. Both the text-only (5.875 errors) and image+text (5.400 errors) baselines achieve comparable performance, with image+text showing a slight edge. Our proposed 3-phase framework yields a mean error of 6.125, which is marginally higher than the image+text baseline but close to the text-only baseline.

Analyzing individual document performance, our proposed method achieved the lowest error count in 17 out of 40 test samples, while the image+text baseline led in 13 documents and text-only in 15. This indicates that, with InternVL3, the benefit of our zero-shot schema-driven approach is comparable to a straightforward image+text prompt. However, it does not consistently outperform the image+text baseline for every document, particularly struggling with forms requiring precise OCR driven table parsing (like invoices and insurance documents). This suggests that Phase 1 segmentation can be hindered if InternVL3's underlying vision language encoder doesn't reliably detect block boundaries in such complex layouts.

In contrast, when Gemma3 serves as the MLLM (Table 7.7), we observe a clear advantage for our proposed framework. The image+text baseline achieves a mean of 4.725 errors, while our proposed method significantly reduces this to 3.275 errors—a 30 percent improvement. Compared to the text-only (5.950 errors) and image-only (13.525 errors) baselines, the proposed method is substantially better across the board.

Proposed + Gemma3 not only has the lowest mean error (3.275), but it also performs best in 32 of 40 documents, outperforming image+text and text-only baselines. This clearly shows that, when combined with a stronger underlying vision language model, the 3-phase framework consistently produces more accurate extractions. The dynamic schema generation and verification steps work well with Gemma3's high quality features, resulting in reliable performance across all document types.

Table 7.8 compares the error counts and extraction issues in sample `invoice 5`. The image-only baseline has 22 errors, misreading important fields like account numbers, advertiser and agency identifiers, and line item details. Although the text-only baseline reduces errors to 7, it still encounters issues with missing dollar signs, merged AD-ID fields, and concatenated table headers that limit accurate parsing. When both an image and text are provided, errors decrease to 9 as some OCR inconsistencies are corrected, but header splitting and truncated program names remain. Our 3-phase framework reduces errors to two, resulting in minor street number errors and a single misrecorded line item amount. The schema enforcement and LLM based verification effectively address extraction issues that simpler baselines failed to.

7.5 Discussion

While our 3-phase zero-shot framework offers a flexible, schema driven approach to multimodal document understanding, it is not without limitations. The framework's effectiveness is primarily influenced by the capabilities of the underlying vision language model and the inherent challenges of real-world document processing. As observed in our experiments, the framework relies heavily on the MLLM's ability to accurately segment documents into coherent information blocks, recognize key value relationships, and precisely follow prompt instructions. Notably, when using a less capable MLLM such as InternVL3, errors become significantly more frequent, particularly with documents that feature complex layouts.

Another critical requirement is OCR accuracy. If OCR fails to correctly recognize critical fields, the extracted JSON will inevitably contain errors. Furthermore, documents with highly irregular or deeply nested layouts can complicate the extraction logic, resulting in misplaced or overlapping information. While Phase 3's LLM based verification can detect some of these errors, it cannot recover data that was never correctly captured due to faulty segmentation or initial OCR errors.

Our evaluation utilized 40 real world documents across four categories: resumes, insurance forms, invoices, and tickets. Although this selection covers a variety of layouts, it doesn't encompass the full spectrum of document types found in enterprise environments. Therefore, the framework's generalization to entirely novel document categories remains an area for further investigation. In production settings where new document categories emerge frequently, maintaining and validating these prompt templates could also become labor intensive.

Additionally, since Phase 3 utilizes the same LLM for self review, the model's inherent biases and tendencies towards hallucination remain a risk. If the LLM consistently misinterprets a block in the same way during both extraction and verification, Phase 3 may fail to correct the mistake.

For very sparse documents (e.g., tickets with minimal fields), the overhead of block segmentation can sometimes introduce unknown structure blocks, slightly increasing false positives compared to a straightforward image+text baseline. Moreover, the 3-phase pipeline incurs additional LLM calls (one per phase), which can increase inference time by up to three times compared to a single unified prompt. This trade-off between accuracy and speed needs careful consideration in real-time or high volume deployment scenarios.

7.6 Conclusion

In summary, this study presents a 3-phase, zero-shot information extraction framework that leverages the capabilities of multimodal large language models to advance document understanding, particularly for visually complex documents. The framework's phased design, which includes document understanding and block identification, schema driven data extraction, and LLM based verification, enables dynamic schema generation and incorporates effective self correction mechanisms. This structure allows the framework to adapt to a wide range of document types without requiring explicit, template specific training.

The effectiveness of the framework is validated through experimental evaluations with InternVL3 and Gemma3 as MLLM. While the framework's performance with InternVL3 is consistent with naive prompting strategies, making use of Gemma3 results in a 30 percent reduction in mean extraction errors. This emphasizes the critical role of strong visual text alignment in achieving high accuracy and robustness across various layouts.

However, several limitations remain, including the framework's dependence on precise OCR, its capacity to manage documents with irregular structures, the complex nature of prompting, and the computational requirements of the multi phase pipeline. Future research should investigate strategies to address these limitations, including the implementation of secondary verification modules to improve accuracy, the refinement of block type taxonomies for better document comprehension, and the expansion of support to include a wider variety of document formats and types.

Overall, this study provides a reliable and adaptable solution to the problem of zero-shot document information extraction. By focusing on structured processing, dynamic schema generation, and optional verification, the framework provides a scalable path to more intelligent, automated document processing systems. Its zero-shot nature allows for rapid adaptation across domains, resulting in extraction performance suitable for real-world production environments.

8 Conclusion

> *You have survived all of your worst days, but eventually one of them is going to get you.*
>
> **— Dave Tarnowski**

This thesis advances and challenges document understanding, particularly layout analysis, table detection, and classification, which are crucial to structural information extraction. The research improved document analysis accuracy and robustness, especially in complex layout identification, table recognition, and document classification, using deep learning. This work also provided new datasets, model architectures, and strategies to fill gaps and improve data handling in these processes.

8.1 Summary of Contributions

Chapter 2 demonstrated the efficacy of object detection principles for document layout analysis, highlighting the potential of CNN architectures in this context. On the PubLayNet dataset, our approach outperformed the baseline methods by 3 percent. Chapter 3 successfully mitigated catastrophic forgetting in table detection by implementing continual learning techniques, particularly experience replay, which resulted in a 15 percent reduction in the forgetting effect compared to traditional fine-tuning methods. Chapter 4 introduced a novel CNN model and a custom dataset, TabLines, for ruling line segmentation, and demonstrated the method's effectiveness on the TabLines dataset.

Furthermore, Chapter 5 tackled the issue of imbalanced document classification through a multi-modal learning approach, integrating image and text data for enhanced accuracy. The visual stream utilized ResNet, employing a two-phase classification strategy with cross-entropy loss and IB loss function to mitigate imbalance effects. The combination of outputs from ResNet and BERT, along with techniques to address class imbalance, resulted in improved performance.

Finally, this thesis looked into the potential of Large Language Models for advanced information extraction. Chapter 6 presented a novel approach for fine-tuning LLMs to extract meaningful structural information from visually rich documents. By converting OCR outputs to structured HTML representations and using instruction-based prompting, we improved LLMs' ability to understand and extract layout-dependent information. Our fine-tuned CodeLlama model outperformed baselines on the VRDU benchmark by more than 20 percent, yielding results comparable to the SOTA.

Chapter 7 builds on this by presenting a three-phase, zero-shot information extraction framework for complex documents that employs Multimodal LLMs. This phased design (document understanding, schema-driven extraction, and LLM-based verification) allows for dynamic schema generation and self-correction. Evaluations revealed a 30 percent reduction in mean extraction errors with Gemma3 over naive prompting, emphasizing the importance of strong visual-text alignment and providing a scalable path to intelligent, automated zero-shot document processing.

The study's findings have important implications for document understanding. Improved document analysis accuracy and robustness may assist in the automation of document processing tasks across a wide range of domains. Notably, this research provides techniques for dealing with real-world data challenges such as data variability, catastrophic forgetting, noise and scarcity, and class imbalance, as well as enabling generalization with few or no task-specific examples (few-shot and zero-shot learning). Furthermore, the study of LLMs and MLLMs in document understanding opens up new research opportunities for leveraging these powerful models to extract meaningful information from complex documents. Overall, this study has explored important challenges in document comprehension, laying the groundwork for more dependable, efficient, and adaptable future research and applications in this field.

Bibliography

[1] Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Document layout analysis with an enhanced object detector”. In: *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE. 2021, pp. 1–5 (cit. on pp. 5, 135).

[2] Mohammad Minouei, Khurram Azeem Hashmi, Mohammad Reza Soheili, Muhammad Zeshan Afzal, and Didier Stricker. “Continual learning for table detection in document images”. In: *Applied Sciences* 12.18 (2022), p. 8969 (cit. on pp. 5, 135).

[3] Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Efficient table border segmentation with asymmetric convolutions”. In: *Fourteenth International Conference on Machine Vision (ICMV 2021)*. Vol. 12084. SPIE. 2022, pp. 133–140 (cit. on pp. 5, 135).

[4] Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Multi-Modal Approach for Imbalanced Document Classification”. In: *17th International Conference on Machine Vision (ICMV 2024)*. SPIE. 2024, pp. 133–140 (cit. on pp. 5, 135).

[5] Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Embedding Layout in Text for Document Understanding Using Large Language Models”. In: *International Conference on Document Analysis and Recognition*. Springer. 2024, pp. 280–293 (cit. on pp. 6, 135).

[6] Jwalin Bhatt, Khurram Azeem Hashmi, Muhammad Zeshan Afzal, and Didier Stricker. “A Survey of Graphical Page Object Detection with Deep Neural Networks”. In: *Applied Sciences* 11.12 (2021), p. 5344 (cit. on pp. 7, 10, 23, 29).

[7] Lawrence O’Gorman. “The document spectrum for page layout analysis”. In: *IEEE Transactions on pattern analysis and machine intelligence* 15.11 (1993), pp. 1162–1173 (cit. on p. 8).

[8] Koichi Kise, Osamu Yanagida, and Shinobu Takamatsu. “Page segmentation based on thinning of background”. In: *Proceedings of 13th International Conference on Pattern Recognition*. Vol. 3. IEEE. 1996, pp. 788–792 (cit. on p. 8).

[9] Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl. “Document analysis system”. In: *IBM journal of research and development* 26.6 (1982), pp. 647–656 (cit. on p. 8).

[10] Oleg Okun, David Dœrmann, and Matti Pietikainen. *Page segmentation and zone classification: the state of the art*. Tech. rep. OULU UNIV (FINLAND) DEPT OF ELECTRICAL ENGINEERING, 1999 (cit. on p. 8).

[11]Jaekyu Ha, R.M. Haralick, and I.T. Phillips. “Recursive X-Y cut using bounding boxes of connected components”. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 2. 1995, 952–955 vol.2 (cit. on p. 8).

[12]Galal M. Binmakhashen and Sabri A. Mahmoud. “Document Layout Analysis: A Comprehensive Survey”. In: *ACM Comput. Surv.* 52.6 (Oct. 2019) (cit. on p. 9).

[13]Sébastien Eskenazi, Petra Gomez-Krämer, and Jean-Marc Ogier. “A comprehensive survey of mostly textual document segmentation algorithms since 2008”. In: *Pattern Recognition* 64 (2017), pp. 1–14 (cit. on p. 9).

[14]Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. “A Survey of Deep Learning Approaches for OCR and Document Understanding”. In: *arXiv preprint arXiv:2011.13534* (2020) (cit. on p. 10).

[15]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on p. 10).

[16]Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149 (cit. on p. 10).

[17]Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection”. In: *Proc. IEEE Conf. Comput. vision pattern Recognit.* 2016, pp. 779–788 (cit. on pp. 10, 11).

[18]Licheng Jiao, Fan Zhang, Fang Liu, et al. “A survey of deep learning-based object detection”. In: *IEEE Access* 7 (2019), pp. 128837–128868 (cit. on pp. 10, 11).

[19]Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125 (cit. on p. 10).

[20]X. Yi, L. Gao, Y. Liao, et al. “CNN Based Page Object Detection in Document Images”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 230–235 (cit. on p. 11).

[21]S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed. “DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 1162–1167 (cit. on p. 12).

[22]Ranajit Saha, Ajoy Mondal, and CV Jawahar. “Graphical object detection in document images”. In: *Int. Conf. Document Anal. Recognit. (ICDAR)*. Sydney, Australia, September 20–25, 2019, pp. 51–58 (cit. on pp. 12, 23).

[23]Ankur Goswami, Joshua McGrath, Shanan Peters, and Theodoros Rekatsinas. “Fine-Grained Object Detection over Scientific Document Images with Region Embeddings”. In: *arXiv preprint arXiv:1910.12462* (2019) (cit. on p. 12).

[24]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916 (cit. on pp. 12, 14).

[25]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 13, 57).

[26]Hang Zhang, Chongruo Wu, Zhongyue Zhang, et al. “Resnest: Split-attention networks”. In: *arXiv preprint arXiv:2004.08955* (2020) (cit. on p. 13).

[27]Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. “Aggregated residual transformations for deep neural networks”. In: *Proc. IEEE Conf. Comput. vision pattern Recognit.* 2017, pp. 1492–1500 (cit. on p. 13).

[28]Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141 (cit. on p. 13).

[29]Thang Vu, Hyunjun Jang, Trung X Pham, and Chang D Yoo. “Cascade rpn: Delving into high-quality region proposal network with adaptive convolution”. In: 2019 (cit. on p. 14).

[30]Kai Chen, Jiaqi Wang, Jiangmiao Pang, et al. “Mmdetection: Open mmlab detection toolbox and benchmark”. In: *arXiv preprint arXiv:1906.07155* (2019) (cit. on p. 15).

[31]Chao Peng, Tete Xiao, Zeming Li, et al. “Megdet: A large mini-batch object detector”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6181–6189 (cit. on p. 16).

[32]Apostolos Antonacopoulos, Basilis Gatos, and Dimosthenis Karatzas. “ICDAR2003 page segmentation competition”. In: (2003) (cit. on p. 16).

[33]Apostolos Antonacopoulos, Stefan Pletschacher, David Bridson, and Christos Papadopoulos. “ICDAR 2009 page segmentation competition”. In: *2009 10th International Conference on Document Analysis and Recognition*. IEEE. 2009, pp. 1370–1374 (cit. on p. 16).

[34]Apostolos Antonacopoulos, Christian Clausner, Christos Papadopoulos, and Stefan Pletschacher. “ICDAR2015 competition on recognition of documents with complex layouts-RDCL2015”. In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2015, pp. 1151–1155 (cit. on p. 16).

[35]L. Gao, X. Yi, Z. Jiang, L. Hao, and Z. Tang. “ICDAR2017 Competition on Page Object Detection”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 1417–1422 (cit. on p. 16).

[36]Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. “Publaynet: largest dataset ever for document layout analysis”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1015–1022 (cit. on pp. 16, 18, 20).

[37]*COCO - Common Objects in Context* (cit. on p. 18).

[38]Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. “Mask r-cnn”. In: *Proc. IEEE Int. Conf. Comput. vision*. 2017, pp. 2961–2969 (cit. on pp. 18, 23).

[39]Xiao-Hui Li, Fei Yin, and Cheng-Lin Liu. “Page Segmentation Using Convolutional Neural Network and Graphical Model”. In: *International Workshop on Document Analysis Systems*. Springer. 2020, pp. 231–245 (cit. on p. 20).

[40]Robert M French. “Catastrophic forgetting in connectionist networks”. In: *Trends in cognitive sciences* 3.4 (1999), pp. 128–135 (cit. on pp. 21, 24).

[41]Katsuhiko Itonori. “Table structure recognition based on textblock arrangement and ruled line position”. In: *Proc. 2nd Int. Conf. Document Anal. Recognit. (ICDAR'93)*. Tsukuba City, Japan, October 20-22, 1993, pp. 765–768 (cit. on p. 22).

[42]Surekha Chandran and Rangachar Kasturi. “Structural recognition of tabulated data”. In: *Proc. 2nd Int. Conf. Document Anal. Recognit. (ICDAR'93)*. sukuba, Japan, October 20–22, 1993, pp. 516–519 (cit. on p. 22).

[43]Pallavi Pyreddy and W Bruce Croft. “Tintin: A system for retrieval in text tables”. In: *Proc. 2nd ACM Int. Conf. Digit. libraries*. 1997, pp. 193–200 (cit. on p. 22).

[44]E Green and M Krishnamoorthy. “Recognition of tables using table grammars”. In: *Proc. 4th Annu. Symp. Document Anal. Inf. Retrieval*. 1995, pp. 261–278 (cit. on p. 22).

[45]B Coüasnon and A Lemaitre. “Handbook of Document Image Processing and Recognition, chapter Recognition of Tables and Forms”. In: *D. Doermann and K. Tombre, Eds. London, U.K.: Springer* (2014), pp. 647–677 (cit. on p. 22).

[46]David W Embley, Matthew Hurst, Daniel Lopresti, and George Nagy. “Table-processing paradigms: a research survey”. In: *Int. J. Document Anal. Recognit. (IJDAR)* 8.2-3 (2006), pp. 66–86 (cit. on p. 22).

[47]Richard Zanibbi, Dorothea Blostein, and James R Cordy. “A survey of table recognition”. In: *Document Anal. Recognit.* 7.1 (2004), pp. 1–16 (cit. on p. 23).

[48]Ana Costa e Silva, Alípio M Jorge, and Luís Torgo. “Design of an end-to-end method to extract information from tables”. In: *Int. J. Document Anal. Recognit. (IJDAR)* 8.2-3 (2006), pp. 144–171 (cit. on p. 23).

[49]Shah Khusro, Asima Latif, and Irfan Ullah. “On methods and tools of table detection, extraction and annotation in PDF documents”. In: *J. Inf. Sci.* 41.1 (2015), pp. 41–57 (cit. on p. 23).

[50]Leipeng Hao, Liangcai Gao, Xiaohan Yi, and Zhi Tang. “A table detection method for pdf documents based on convolutional neural networks”. In: *12th IAPR Workshop Document Anal. Sys. (DAS)*. Santorini, Greece, April 11–14, 2016, pp. 287–292 (cit. on p. 23).

[51]Azka Gilani, Shah Rukh Qasim, Imran Malik, and Faisal Shafait. “Table detection using deep learning”. In: *14th IAPR Int. Conf. document Anal. Recognit. (ICDAR)*. Vol. 1. Kyoto, Japan, November 9–15, 2017, pp. 771–776 (cit. on p. 23).

[52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *arXiv:1506.01497* (2015) (cit. on pp. 23, 29).

[53] Shoaib Ahmed Siddiqui, Muhammad Imran Malik, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. “Decnt: Deep deformable cnn for table detection”. In: *IEEE Access* 6 (2018), pp. 74151–74161 (cit. on p. 23).

[54] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. “CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents”. In: *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*. Seattle, WA, USA, June 14–19, 2020, pp. 572–573 (cit. on p. 23).

[55] Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. “CasTabDetectoRS: Cascade Network for Table Detection in Document Images with Recursive Feature Pyramid and Switchable Atrous Convolution”. In: *Journal of Imaging* 7.10 (2021), p. 214 (cit. on pp. 23, 38).

[56] Khurram Azeem Hashmi, Didier Stricker, Marcus Liwicki, Muhammad Noman Afzal, and Muhammad Zeshan Afzal. “Guided Table Structure Recognition through Anchor Optimization”. In: *arXiv:2104.10538* (2021) (cit. on pp. 23, 42).

[57] Danish Nazir, Khurram Azeem Hashmi, Alain Pagani, et al. “HybridTabNet: Towards better table detection in scanned document images”. In: *Applied Sciences* 11.18 (2021), p. 8396 (cit. on p. 23).

[58] Kai Chen, Jiangmiao Pang, Jiaqi Wang, et al. “Hybrid task cascade for instance segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4974–4983 (cit. on p. 23).

[59] Isaak Kavasidis, Sergio Palazzo, Concetto Spampinato, et al. “A saliency-based convolutional neural network for table and chart detection in digitized documents”. In: *arXiv:1804.06236* (2018) (cit. on p. 23).

[60] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. “Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images”. In: *Int. Conf. Document Anal. Recognit. (ICDAR)*. Sydney, Australia, September 20–25, 2019, pp. 128–133 (cit. on p. 23).

[61] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. “Rethinking table recognition using graph neural networks”. In: *Int. Conf. Document Anal. Recognit. (ICDAR)*. Sydney, Australia, September 20–25, 2019, pp. 142–147 (cit. on p. 23).

[62] Martin Holeček, Antonín Hoskovec, Petr Baudiš, and Pavel Klinger. “Table understanding in structured documents”. In: *Int. Conf. Document Anal. Recognit. Workshops (ICDARW)*. Vol. 5. Sydney, Australia, September 20–25, 2019, pp. 158–164 (cit. on p. 23).

[63] Yilun Huang, Qinqin Yan, Yibo Li, et al. “A YOLO-based table detection method”. In: *Int. Conf. Document Anal. Recognit. (ICDAR)*. Sydney, Australia, September 20–25, 2019, pp. 813–818 (cit. on p. 23).

[64] Madhav Agarwal, Ajoy Mondal, and CV Jawahar. “CDeC-Net: Composite Deformable Cascade Network for Table Detection in Document Images”. In: *arXiv:2008.10831* (2020) (cit. on pp. 23, 38).

[65] Saman Arif and Faisal Shafait. “Table detection in document images using foreground and background features”. In: *Digit. Image Computing : Techn. Appl. (DICTA)*. 2018, pp. 1–8 (cit. on p. 23).

[66] Ningning Sun, Yuanping Zhu, and Xiaoming Hu. “Faster R-CNN based table detection combining corner locating”. In: *Int. Conf. Document Anal. Recognit. (ICDAR)*. Sydney, Australia, September 20–25, 2019, pp. 1314–1319 (cit. on p. 23).

[67] Khurram Azeem Hashmi, Marcus Liwicki, Didier Stricker, et al. “Current Status and Performance Analysis of Table Recognition in Document Images with Deep Neural Networks”. In: *IEEE Access* (2021) (cit. on pp. 23, 43).

[68] Matthias Delange, Rahaf Aljundi, Marc Masana, et al. “A continual learning survey: Defying forgetting in classification tasks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021) (cit. on p. 24).

[69] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-jussà. “Continual lifelong learning in natural language processing: A survey”. In: *arXiv preprint arXiv:2012.09823* (2020) (cit. on p. 24).

[70] Stephen Grossberg. “How does a brain build a cognitive code?” In: *Studies of mind and brain* (1982), pp. 1–52 (cit. on p. 24).

[71] Haoxuan Qu, Hossein Rahmani, Li Xu, Bryan Williams, and Jun Liu. “Recent Advances of Continual Learning in Computer Vision: An Overview”. In: *arXiv preprint arXiv:2109.11369* (2021) (cit. on p. 24).

[72] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. “Overcoming catastrophic forgetting in neural networks”. In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526 (cit. on p. 24).

[73] Liyang Liu, Zhanghui Kuang, Yimin Chen, et al. “Incdet: In defense of elastic weight consolidation for incremental object detection”. In: *IEEE transactions on neural networks and learning systems* 32.6 (2020), pp. 2306–2319 (cit. on p. 24).

[74] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. “iCaRL: Incremental Classifier and Representation Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017 (cit. on p. 24).

[75] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* 2.7 (2015) (cit. on p. 24).

[76] Junting Zhang, Jie Zhang, Shalini Ghosh, et al. “Class-incremental learning via deep model consolidation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 1131–1140 (cit. on p. 24).

[77] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. “Incremental learning of object detectors without catastrophic forgetting”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3400–3409 (cit. on p. 25).

[78]Ross Girshick. “Fast R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015 (cit. on pp. 25, 29).

[79]C Lawrence Zitnick and Piotr Dollár. “Edge boxes: Locating object proposals from edges”. In: *European conference on computer vision*. Springer. 2014, pp. 391–405 (cit. on p. 25).

[80]Manoj Acharya, Tyler L Hayes, and Christopher Kanan. “Rodeo: Replay for online object detection”. In: *arXiv preprint arXiv:2008.06439* (2020) (cit. on p. 26).

[81]Jeng-Lun Shieh, Muhamad Amirul Haq, Said Karam, et al. “Continual learning strategy in one-stage object detection framework based on experience replay for autonomous driving vehicle”. In: *Sensors* 20.23 (2020), p. 6777 (cit. on p. 26).

[82]Nicolas Carion, Francisco Massa, Gabriel Synnaeve, et al. “End-to-end object detection with transformers”. In: *European conference on computer vision*. Springer. 2020, pp. 213–229 (cit. on p. 29).

[83]Wenhai Wang, Enze Xie, Xiang Li, et al. “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 568–578 (cit. on pp. 29, 30).

[84]Wenhai Wang, Enze Xie, Xiang Li, et al. “Pvtv2: Improved baselines with pyramid vision transformer”. In: *Computational Visual Media* 8.3 (2022), pp. 1–10 (cit. on pp. 29, 30).

[85]Peize Sun, Rufeng Zhang, Yi Jiang, et al. “Sparse r-cnn: End-to-end object detection with learnable proposals”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14454–14463 (cit. on pp. 29, 30).

[86]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 29, 30).

[87]Kai Chen, Jiaqi Wang, Jiangmiao Pang, et al. “MMDetection: Open MMLab Detection Toolbox and Benchmark”. In: *arXiv preprint arXiv:1906.07155* (2019) (cit. on p. 30).

[88]Jia Deng, Wei Dong, Richard Socher, et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 30).

[89]Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, et al. “Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming”. In: *arXiv preprint arXiv:1907.07484* (2019) (cit. on p. 32).

[90]Minghao Li, Lei Cui, Shaohan Huang, et al. “Tablebank: Table benchmark for image-based table detection and recognition”. In: *Proc. The 12th Lang. Resour. Eval. Conf.* Marseille, France, May 11–16, 2020, pp. 1918–1925 (cit. on p. 32).

[91]*arXiv.org e-Print archive*. `https://arxiv.org/`. 2022 (cit. on p. 32).

[92]Brandon Smock, Rohith Pesala, Robin Abraham, and WA Redmond. “PubTables-1M: Towards comprehensive table extraction from unstructured documents”. In: *arXiv preprint arXiv:2110.00061* (2021) (cit. on pp. 32, 38).

[93]Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. “Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context”. In: *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision*. 2021, pp. 697–706 (cit. on p. 33).

[94]Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. “Image-based table recognition: data, model, and evaluation”. In: *arXiv:1911.10683* (2019) (cit. on p. 33).

[95]Richard O. Duda and Peter E. Hart. “Use of the Hough transformation to detect lines and curves in pictures”. In: *Commun. ACM* 15 (1972), pp. 11–15 (cit. on p. 42).

[96]Thotreingam Kasar, Philippine Barlas, Sebastien Adam, Clément Chatelain, and Thierry Paquet. “Learning to detect tables in scanned document images using line information”. In: *12th Int. Conf. Document Anal. Recognit.* Washington, DC, USA, August 25–28, 2013, pp. 1185–1189 (cit. on p. 42).

[97]Hong Tai Tran, Tuan Anh Tran, In Seop Na, and Soo Hyung Kim. “Cell decomposition for the table in document image based on analysis of texts and lines distribution”. In: *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN)*. 2016, pp. 736–738 (cit. on p. 42).

[98]Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. “DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 1162–1167 (cit. on p. 42).

[99]Shoaib Ahmed Siddiqui, Imran Ali Fateh, Syed Tahseen Raza Rizvi, Andreas Dengel, and Sheraz Ahmed. “Deeptabstr: Deep learning based table structure recognition”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1403–1409 (cit. on p. 42).

[100]Sachin Raja, Ajoy Mondal, and CV Jawahar. “Table Structure Recognition using Top-Down and Bottom-Up Cues”. In: *Eur. Conf. Comput. Vision*. Springer, Cham. 2020, pp. 70–86 (cit. on p. 42).

[101]D. Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita A. Sultanpure. “CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), pp. 2439–2447 (cit. on pp. 42, 43, 47, 48).

[102]Shoaib Ahmed Siddiqui, Pervaiz Iqbal Khan, Andreas Dengel, and Sheraz Ahmed. “Rethinking Semantic Segmentation for Table Structure Recognition in Documents”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 1397–1402 (cit. on p. 42).

[103] Yajun Zou and Jinwen Ma. "A Deep Semantic Segmentation Model for Image-based Table Structure Recognition". In: *15th IEEE Int. Conf. Signal Process. (ICSP)*. Vol. 1. 2020, pp. 274–280 (cit. on p. 42).

[104] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, and Jing-Jhih Lin. "Efficient dense modules of asymmetric convolution for real-time semantic segmentation". In: *Proceedings of the ACM Multimedia Asia*. 2019, pp. 1–6 (cit. on pp. 43, 44, 48).

[105] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495 (cit. on p. 43).

[106] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Int. Conf. Med. image Comput. Comp. intervention*. Springer, Cham. 2015, pp. 234–241 (cit. on p. 43).

[107] Christian Szegedy, Wei Liu, Yangqing Jia, et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9 (cit. on p. 43).

[108] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. on p. 44).

[109] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. "Large kernel matters–improve semantic segmentation by global convolutional network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4353–4361 (cit. on p. 44).

[110] Alex Bäuerle, Christian Van Onzenoodt, and Timo Ropinski. "Net2Vis–A Visual Grammar for Automatically Generating Publication-Tailored CNN Architecture Visualizations". In: *IEEE transactions on visualization and computer graphics* 27.6 (2021), pp. 2980–2991 (cit. on p. 45).

[111] Martín Abadi, Ashish Agarwal, Paul Barham, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015 (cit. on p. 45).

[112] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456 (cit. on p. 45).

[113] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034 (cit. on p. 45).

[114] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988 (cit. on pp. 46, 55).

[115]Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. "Extracting scientific figures with distantly supervised neural networks". In: *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*. 2018, pp. 223–232 (cit. on p. 46).

[116]Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. "ICDAR 2013 table competition". In: *2013 12th International Conference on Document Analysis and Recognition*. IEEE. 2013, pp. 1449–1453 (cit. on p. 46).

[117]Zewen Chi, Heyan Huang, Heng-Da Xu, et al. "Complicated table structure recognition". In: *arXiv preprint arXiv:1908.04729* (2019) (cit. on p. 46).

[118]Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. "Image-based table recognition: data, model, and evaluation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer. 2020, pp. 564–580 (cit. on p. 46).

[119]Alexander B. Jung, Kentaro Wada, Jon Crall, et al. *imgaug*. `https://github.com/aleju/imgaug`. Online; accessed 01-Feb-2020. 2020 (cit. on pp. 46, 47).

[120]Chaowei Fang, Dingwen Zhang, Wen Zheng, et al. "Revisiting Long-tailed Image Classification: Survey and Benchmarks with New Evaluation Metrics". In: *arXiv preprint arXiv:2302.01507* (2023) (cit. on p. 54).

[121]Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural networks* 106 (2018), pp. 249–259 (cit. on p. 54).

[122]Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. "Parametric contrastive learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 715–724 (cit. on p. 54).

[123]Yijing Li, Haixiang Guo, Qingpeng Zhang, Mingyun Gu, and Jianying Yang. "Imbalanced text sentiment classification using universal and domain-specific knowledge". In: *Knowledge-Based Systems* 160 (2018), pp. 1–15 (cit. on p. 55).

[124]Xinyi Gao, Wentao Zhang, Tong Chen, et al. "Semantic-aware Node Synthesis for Imbalanced Heterogeneous Information Networks". In: *arXiv preprint arXiv:2302.14061* (2023) (cit. on p. 55).

[125]Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. "Long-tailed recognition by routing diverse distribution-aware experts". In: *arXiv preprint arXiv:2010.01809* (2020) (cit. on p. 55).

[126]Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. "Influence-balanced loss for imbalanced visual classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 735–744 (cit. on p. 55).

[127]Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on p. 57).

[128] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114 (cit. on p. 57).

[129] Mindee. *docTR: Document Text Recognition*. `https://github.com/mindee/doctr`. 2021 (cit. on p. 58).

[130] Adam Paszke, Sam Gross, Francisco Massa, et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: `1912.01703 [cs.LG]` (cit. on p. 59).

[131] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017) (cit. on p. 60).

[132] Yang Xu, Yiheng Xu, Tengchao Lv, et al. "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding". In: *arXiv preprint arXiv:2012.14740* (2020) (cit. on pp. 61, 72, 85).

[133] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. "Docformer: End-to-end transformer for document understanding". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 993–1003 (cit. on p. 61).

[134] Jaekyu Ha, R.M. Haralick, and I.T. Phillips. "Recursive X-Y cut using bounding boxes of connected components". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 2. 1995, 952–955 vol.2 (cit. on p. 71).

[135] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. "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 (cit. on pp. 71, 72, 85, 88).

[136] Humza Naveed, Asad Ullah Khan, Shi Qiu, et al. "A comprehensive overview of large language models". In: *arXiv preprint arXiv:2307.06435* (2023) (cit. on pp. 71, 74).

[137] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023) (cit. on pp. 71, 73).

[138] Hugo Touvron, Louis Martin, Kevin Stone, et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023) (cit. on pp. 71, 73, 77).

[139] Yiheng Xu, Minghao Li, Lei Cui, et al. "Layoutlm: Pre-training of text and layout for document image understanding". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1192–1200 (cit. on pp. 72, 85).

[140] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 72).

[141] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, et al. “Formnet: Structural encoding beyond sequential modeling in form document information extraction”. In: *arXiv preprint arXiv:2203.08411* (2022) (cit. on pp. 72, 85).

[142] Geewook Kim, Teakgyu Hong, Moonbin Yim, et al. “Ocr-free document understanding transformer”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 498–517 (cit. on pp. 72, 88).

[143] Vincent Perot, Kai Kang, Florian Luisier, et al. “LMDX: Language Model-based Document Information Extraction and Localization”. In: *arXiv preprint arXiv:2309.10952* (2023) (cit. on pp. 72, 85, 86, 88).

[144] A Waswani, N Shazeer, N Parmar, et al. “Attention is all you need”. In: *NIPS*. 2017 (cit. on p. 73).

[145] Tom Brown, Benjamin Mann, Nick Ryder, et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on p. 74).

[146] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, et al. “Code llama: Open foundation models for code”. In: *arXiv preprint arXiv:2308.12950* (2023) (cit. on pp. 77, 87).

[147] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, et al. *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods*. `https://github.com/huggingface/peft`. 2022 (cit. on p. 77).

[148] Edward J Hu, Yelong Shen, Phillip Wallis, et al. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021) (cit. on p. 77).

[149] *Tesseract Open Source OCR Engine* (cit. on pp. 79, 103).

[150] Thomas M Breuel. “The hOCR microformat for OCR workflow and results”. In: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 1063–1067 (cit. on p. 79).

[151] Izzeddin Gur, Ofir Nachum, Yingjie Miao, et al. “Understanding html with large language models”. In: *arXiv preprint arXiv:2210.03945* (2022) (cit. on p. 79).

[152] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. “Qlora: Efficient finetuning of quantized llms”. In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on p. 82).

[153] Seunghyun Park, Seung Shin, Bado Lee, et al. “CORD: a consolidated receipt dataset for post-OCR parsing”. In: *Workshop on Document Intelligence at NeurIPS 2019*. 2019 (cit. on pp. 83, 84).

[154] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. “Vrdu: A benchmark for visually-rich document understanding”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 5184–5193 (cit. on p. 84).

[155] *DeciAI Research Team. 2024. DeciLM-7B-instruct* (cit. on pp. 87, 88).

[156]Davide Caffagni, Federico Cocchi, Luca Barsellotti, et al. “The revolution of multimodal large language models: a survey”. In: *arXiv preprint arXiv:2402.12451* (2024) (cit. on pp. 91–93).

[157]Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, et al. “A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision-Language Tasks”. In: *arXiv preprint arXiv:2411.06284* (2024) (cit. on pp. 91–93).

[158]Lun-Chi Chen, Hsin-Tzu Weng, Mayuresh Sunil Pardeshi, et al. “Evaluation of Prompt Engineering on the Performance of a Large Language Model in Document Information Extraction”. In: *Electronics* 14.11 (2025), p. 2145 (cit. on pp. 92, 93).

[159]S Yin, C Fu, S Zhao, et al. “A survey on multimodal large language models. arXiv 2023”. In: *arXiv preprint arXiv:2306.13549* () (cit. on pp. 93, 103).

[160]Zijing Liang, Yanjie Xu, Yifan Hong, et al. “A Survey of Multimodel Large Language Models”. In: *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*. 2024, pp. 405–409 (cit. on p. 93).

[161]Aniket Bhattacharyya, Anurag Tripathi, Ujjal Das, et al. “Information Extraction from Visually Rich Documents using LLM-based Organization of Documents into Independent Textual Segments”. In: *arXiv preprint arXiv:2505.13535* (2025) (cit. on p. 93).

[162]Ye Mo, Zirui Shao, Kai Ye, et al. “Doc-CoB: Enhancing Multi-Modal Document Understanding with Visual Chain-of-Boxes Reasoning”. In: *arXiv preprint arXiv:2505.18603* (2025) (cit. on p. 94).

[163]Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. “Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models”. In: *arXiv preprint arXiv:2504.10479* (2025) (cit. on pp. 100, 103).

[164]Gemma Team, Aishwarya Kamath, Johan Ferret, et al. “Gemma 3 technical report”. In: *arXiv preprint arXiv:2503.19786* (2025) (cit. on p. 100).

[165]Gemma Team, Morgane Riviere, Shreya Pathak, et al. “Gemma 2: Improving open language models at a practical size”. In: *arXiv preprint arXiv:2408.00118* (2024) (cit. on p. 100).

[166]Aaron Hurst, Adam Lerer, Adam P Goucher, et al. “Gpt-4o system card”. In: *arXiv preprint arXiv:2410.21276* (2024) (cit. on p. 103).

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Mohammad Minouei *Machine Learning Researcher*

linkedin.com/in/matinminouei

PhD in Artificial Intelligence with expertise in deep learning, large language models, and document understanding. Skilled in Python and Java with a strong background in software development, databases, and scalable system design.

Experience

Machine Learning Researcher, *Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)*
09/2020 – 05/2025 | Kaiserslautern, Germany

- Developed custom deep learning architecture for improved object detection and classification.
- Utilized large language models for document information extraction and structural analysis.
- Applied fine-tuning strategies (e.g., LoRA, QLORA) and advanced prompt engineering.
- Reviewed literature and published original research in ICDAR and Applied Sciences.

Software Developer, *Vasni*
06/2014 – 05/2020

- Designed and developed software for desktop and Android devices.
- Built and optimized databases and table structures for web applications.
- Created backend services using RESTful API designs.
- Implemented containerization using Docker, improving application scalability and manageability.

Languages

English ●●●●○
C1

German ●○○○○
A1

Education

PhD Technical University of Kaiserslautern(RPTU), Artificial Intelligence
2020 – 2025 | Kaiserslautern, Germany
Thesis: Structural Information Extraction from Document Images: Addressing Challenges in Layout Analysis, Table Detection, and Classification

M.Sc. Kharazmi University, Artificial Intelligence
2016 – 2019
Thesis: Reconstruction of High-Quality Document Images from Video Sequences

B.Sc. Sadjad University, Software Engineering
2010 – 2015
Thesis: IVR System with Control Panel

Skills

Programming
Python, Java, C++

Deep Learning Frameworks
PyTorch, Keras, Transformers, OpenCV

Environments
Linux, Docker

Databases
SQL, No-SQL

Project Management
Git, and familliar with DevOps

List of Publications

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Document layout analysis with an enhanced object detector”. In: *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE. 2021, pp. 1–5

Mohammad Minouei, Khurram Azeem Hashmi, Mohammad Reza Soheili, Muhammad Zeshan Afzal, and Didier Stricker. “Continual learning for table detection in document images”. In: *Applied Sciences* 12.18 (2022), p. 8969

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Efficient table border segmentation with asymmetric convolutions”. In: *Fourteenth International Conference on Machine Vision (ICMV 2021)*. Vol. 12084. SPIE. 2022, pp. 133–140

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Multi-Modal Approach for Imbalanced Document Classification”. In: *17th International Conference on Machine Vision (ICMV 2024)*. SPIE. 2024, pp. 133–140

Mohammad Minouei, Mohammad Reza Soheili, and Didier Stricker. “Embedding Layout in Text for Document Understanding Using Large Language Models”. In: *International Conference on Document Analysis and Recognition*. Springer. 2024, pp. 280–293