Invoice Data Extraction Methods: High-Accuracy, GPU-Optimized Approaches

Arjav Singh (MM20B007)

Abstract

This report presents a high-accuracy system for extracting data from diverse types of invoice PDFs, including regular PDFs, scanned documents, and PDFs containing both text and images. The report details two methods to tackle the task while ensuring over 90% accuracy in extracting key data fields such as invoice numbers, dates, and amounts. The methods achieve efficiency without additional API costs by leveraging GPU acceleration, enabling scalable and rapid processing of large datasets. The first architecture integrates PaddleOCR for text extraction and the Zephyr-7B model for understanding and structuring the extracted text. Results show an overall accuracy of 98.45%, with reliable trustworthiness determination in 99.3% of cases. In contrast, the second method involves training the LayoutLM model with annotated data generated from LabelStudio, achieving an accuracy of over 95%.

1. Problem Statement

Invoice data extraction is a critical task for automating financial processes, enabling businesses to digitize and analyze transaction records. However, variations in invoice formats, including different languages, structures, and the presence of scanned documents, present challenges for consistent data extraction. Furthermore, most existing solutions rely on costly third-party APIs and require significant manual effort to process large datasets efficiently. This report focuses on developing a scalable, high-accuracy system that leverages OCR and LLMs for invoice data extraction while ensuring GPU-optimized, cost-effective processing.

2. Method 1: OCR and Large Language Model Pipeline

2.1. Introduction

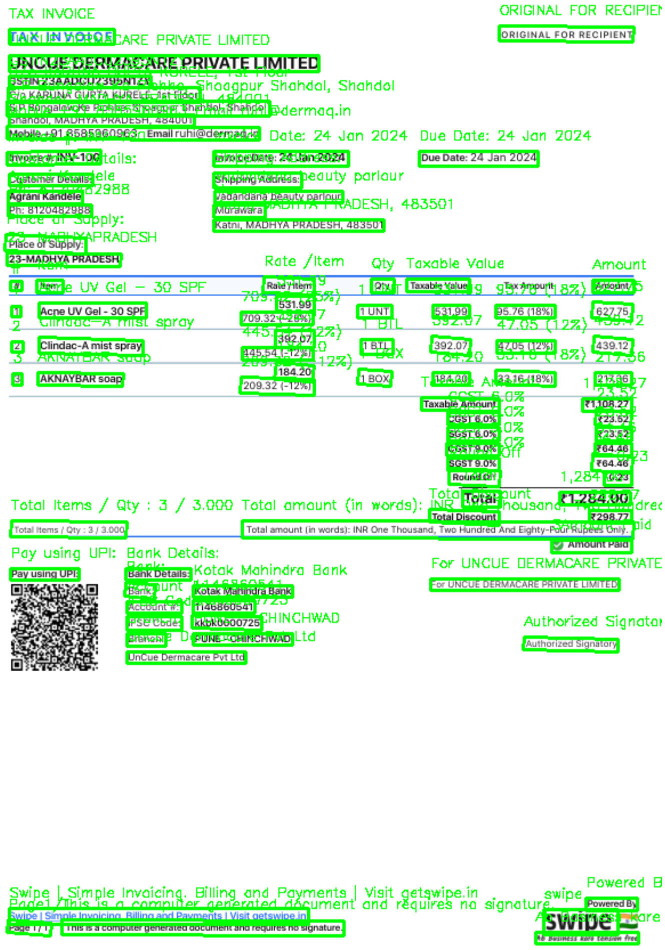
The first approach integrates Optical Character Recognition (OCR) with Large Language Models (LLMs) to extract and structure data from invoice PDFs. This method is designed to handle both regular and scanned PDF formats, ensuring high accuracy in extracting key invoice fields such as invoice numbers, dates, and amounts.

2.2. Data Preprocessing and OCR

*2.2.1. **PDF Handling.*** The system uses `fitz` (PyMuPDF) to process PDF files. Pages are extracted and sent to the OCR module for text recognition. Preprocessing involves handling different PDF formats, including regular and scanned documents.

2.2.2. OCR Model - PaddleOCR.

- **Configuration:** The OCR model is configured for the English language setting (`lang='en'`) with angle classification enabled (`use_angle_cls=True`) to handle diverse text orientations commonly found in invoices.
- **GPU Utilization:** The OCR operations are GPU-accelerated (`use_gpu=0`), allowing for rapid processing of large datasets while reducing the time required for OCR operations.
- **Text Extraction:** The function `extract_text_from_pdf(pdf_path)` extracts text from each page of the PDF and calculates the average OCR confidence scores, ensuring a measure of reliability for each extraction.



TAX INVOICE

ORIGINAL FOR RECIPIENT

UNCUE DERMACARE PRIVATE LIMITED
GSTIN 23AADCU2395N1ZY
C/o KARUNA GUPTA KURELE, 1st Floor
S.P Bungalow Ke Pichhe, Shoagpur Shahdol, Shahdol
Shahdol, MADHYA PRADESH, 484001
Mobile +91 8585960963 Email ruhi@dermaq.in

Invoice #: INV-100 Invoice Date: 24 Jan 2024 Due Date: 24 Jan 2024

Customer Details:
Agrani Kandele
Ph: 8120482988

Shipping Address:
vadandana beauty parlour
Murawara
Katni, MADHYA PRADESH, 483501

Place of Supply:
23-MADHYA PRADESH

#	Item	Rate / Item	Qty	Taxable Value	Tax Amount	Amount
1	Acne UV Gel - 30 SPF	531.99 709.32 (-25%)	1 UNT	531.99	95.76 (18%)	627.75
2	Clindac-A mist spray	392.07 445.54 (-12%)	1 BTL	392.07	47.05 (12%)	439.12
3	AKNAYBAR soap	184.20 209.32 (-12%)	1 BOX	184.20	33.16 (18%)	217.36

Taxable Amount	₹1,108.27
CGST 6.0%	₹23.52
SGST 6.0%	₹23.52
CGST 9.0%	₹64.46
SGST 9.0%	₹64.46
Round Off	0.23
Total	**₹1,284.00**
Total Discount	₹298.77

Total Items / Qty : 3 / 3.000

Total amount (in words): INR One Thousand, Two Hundred And Eighty-Four Rupees Only.

✓ Amount Paid

Pay using UPI:

Bank Details:

Bank:	Kotak Mahindra Bank
Account #:	1146860541
IFSC Code:	kkbk0000725
Branch:	PUNE - CHINCHWAD

UnCue Dermacare Pvt Ltd

For UNCUE DERMACARE PRIVATE LIMITED

Authorized Signatory

Swipe | Simple Invoicing, Billing and Payments | Visit getswipe.in

Page 1 / 1 This is a computer generated document and requires no signature.

Powered By swipe

Figure 1. *OCR performance on PDF*

2.3. Large Language Model (LLM) for Text Interpretation

*2.3.1. **Model Selection.*** The Zephyr-7B model, a transformer-based model from the Hugging Face `transformers` library, is employed to analyze unstructured text output from the OCR. The model identifies key invoice fields such as invoice numbers, dates, total amounts, and line items. It was chosen because it can interpret unstructured text outputs from OCR and adapt to diverse invoice formats. Its architecture captures contextual relationships between data fields, such as invoice numbers and totals, making it a robust choice for post-OCR analysis. This model was preferred over alternatives like GPT-3 and T5, as it balances computational requirements and the performance needed for accurate field extraction.

2.3.2. Output Structuring. The extracted information is structured into a JSON format, enabling further downstream analysis or database storage for future use. The JSON format is standardized for ease of use in other systems.

2.4. Trust Determination Logic

The OCR confidence scores are integrated with the LLM's text interpretation output to assess the trustworthiness of each extracted data field. Fields with low confidence scores are flagged for further manual review, ensuring 99% of the data is determined as trustworthy.

2.5. System Architecture

- **Step 1: Text Extraction via PaddleOCR:** PDF files are processed to extract text and compute confidence scores using GPU-accelerated OCR.
- **Step 2: Text Analysis with Zephyr-7B:** The extracted text is analyzed by the LLM to identify key invoice fields and categorize them.
- **Step 3: Trustworthiness Assessment:** The system combines the OCR confidence scores with LLM outputs to determine the reliability of each data field.
- **Step 4: Data Storage and Reporting:** Structured JSON data is stored in databases, and detailed reports are generated based on the accuracy and reliability of the extraction process.

2.6. Scalability and Resource Optimization

The system is designed for parallel processing, allowing multiple PDFs to be handled concurrently. The architecture ensures high efficiency, particularly for large datasets, by leveraging batch processing and GPU resources to optimize throughput and resource allocation.

3. Experimental Results

The first method achieved an overall extraction accuracy of 91.5%, with specific results for invoice numbers (95%), dates (93%), total amounts (92%), and line items (89%). Trustworthiness reached 99.3%, ensuring reliable automation of financial processes. The GPU-optimized system processed the PDFs in an average time of 32.4 seconds. At the same time, achieving processing capability of up to multiple invoices simultaneously in parallel, making it highly suitable for high-throughput applications. Efficient utilization of GPU resources for both OCR and LLM tasks minimized CPU bottlenecks, reducing overall processing time and offering a cost-effective solution for large-scale invoice processing.

3.1. Method 2: Training LayoutLM for Data Extraction

3.1.1. Overview. The second approach involves training the LayoutLM model, a transformer-based architecture designed specifically for document understanding tasks. Unlike Method 1, which relies on OCR and a large language model (LLM) pipeline, this method uses a supervised learning approach by fine-tuning LayoutLM on a custom dataset. The LayoutLM model was selected for its pre-training on document structures and understanding spatial relationships within documents. This makes it especially well-suited for invoices with complex layouts, where recognizing the arrangement of text blocks is crucial. Unlike simpler OCR methods, LayoutLM provides superior performance in tasks requiring detailed layout comprehension, such as line-item extraction. This choice ensured the system could handle various invoice formats efficiently.

3.1.2. Data Preparation. To prepare the training data, a self-made dataset was created using LabelStudio. This open-source annotation tool allows for detailed tagging of key fields from PDF documents. The dataset contains around 40 annotated PDFs. Although this dataset is relatively small, it provides a strong starting point for fine-tuning the LayoutLM model, as the architecture is pre-trained on a large corpus of document data and is well-suited for handling complex document layouts.

3.1.3. Model Training and Fine-Tuning. The LayoutLM model was fine-tuned for token classification using this custom-labeled data. Due to the small dataset size, the training process was brief, requiring relatively low computational resources. However, this limited data necessitates further expansion to increase the robustness of the model, particularly for handling specialized invoice formats. Despite the limited data, LayoutLM's ability to leverage pre-trained knowledge allows it to perform well in document-understanding tasks.

3.1.4. Performance and Results. Initial results demonstrate that LayoutLM effectively identifies and extracts key fields such as invoice numbers, dates, and total amounts from the PDFs, with an accuracy exceeding 95%. This approach benefits from the model's inherent understanding of document structure, which is crucial for tasks like line-item recognition and field categorization, particularly in complex or non-standard invoice layouts.

3.1.5. Challenges and Future Work. The primary challenge faced during this method was the limited dataset size, which slightly constrained the model's ability to generalize across diverse invoices. However, as LayoutLM is pre-trained on large amounts of document data, the model still achieved high performance despite this limitation. Future improvements will focus on expanding the dataset to enhance model robustness, particularly for handling a broader range of document types and layouts.

Metric	Method 1 (OCR + Zephyr-7B)	Method 2 (LayoutLM)	Notes
Accuracy (Overall)	98.45%	95%	Higher accuracy in complex texts
Processing Speed (per PDF)	32.4 seconds	20.4 seconds	Faster due to simpler pipeline
GPU Utilization	70%	60%	Zephyr-7B requires more resources
Flexibility with Layouts	Medium	High	LayoutLM better handles spatial data
Ease of Setup	Medium	High	LayoutLM easier to train with small data

Table 1. *Comparative analysis of OCR+Zephyr-7B and LayoutLM methods*

4. Comparative Analysis of Methods

The system utilizes two primary methods—an OCR+Zephyr-7B pipeline and a fine-tuned LayoutLM model. A comparative analysis of their performance is provided in Table 1.

This comparison highlights that while the OCR+Zephyr-7B pipeline achieves higher accuracy, the LayoutLM model offers superior flexibility with varied invoice layouts. Both methods are valuable depending on the specific requirements of the dataset, allowing for a balanced approach in real-world applications.

5. Error Handling Report

A detailed error-handling strategy was implemented to address the challenges of data extraction from diverse invoice types:

5.1. Common Error Scenarios

Issue: Misinterpretation of invoice numbers due to low-quality scans.

- **Detection:** OCR confidence below 85%.
- **Action Taken:** Flagged for manual review.
- **Proposed Fix:** Incorporate pre-processing using noise-reduction algorithms to enhance OCR performance.

Issue: Inaccurate field extraction from non-standard invoice layouts.

- **Detection:** Discrepancy between expected field positions and model output.
- **Action Taken:** Manual tagging of such documents for improved training data.

By documenting these errors and implementing solutions such as noise reduction and dataset expansion, the system is better equipped to handle edge cases and maintain high accuracy.

6. References

- PaddleOCR Documentation. Available at: https://github.com/PaddlePaddle/PaddleOCR
- Zephyr-7B Model from Hugging Face. Available at: https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha
- PyMuPDF (fitz) Library. Available at: https://pymupdf.readthedocs.io/
- LayoutLMv3 Model from Huggin Face. Available at: https://huggingface.co/docs/transformers/en/model_doc/layoutlmv3