Document Classification and Information Extraction with Deep Learning techniques

Gunti Lahari
IIT Hyderabad
ai22btech11008@iith.ac.in

J Hima Chandh
IIT Hyderabad
ai22btech11009@iith.ac.in

S Divija
IIT Hyderabad
ai22btech11026@iith.ac.in

K Anuraga Chandan
IIT Hyderabad
ai22btech11011@iith.ac.in

***Abstract*—In the era of Technology driven by data, the demand for the automation process to collect and obtain critical information from any hard-copy document issued by the government is highly in demand. The proposed work is to design a deep learning-based system to extract information from images of Aadhaar cards, PAN cards, and other documents. The drawing out of details from such documents is time-consuming, prone to errors, and expensive when there are massive amounts of documents. Hence, the need for an efficient and accurate system for automating these processes so that mistakes can be avoided and the process rendered smoother.**

The main emphasis of this project is a solution which detects, recognizes, and extracts relevant fields from scanned images of these documents, such as names, identification numbers, addresses, and dates. By doing so, it caters to multiple societal needs, including improving efficiency in the banking sector, simplifying the KYC procedure, and minimizing human errors in form submissions and data entry.

***Keywords*—Classification, Information Extraction, CNN (Convolutional Neural Network), NLP (Natural Language Processing), OCR (Optical Character Recognition)**

I. Introduction

With the increased digitization and expanded use of digital documents, including Aadhaar cards, PAN cards, driving licenses, and voter Ids etc, the need for automated document classification and information extraction has become more critical. These tasks are time-consuming and require a high degree of accuracy. In the era of the growth in the deep learning , the combination of image classification through CNN's and data extraction through the Tesseract Optical Character Recognition (OCR) along with LSTM led to siginificant change in various domains.

Automated document classification (ADC) involves assigning a document to a predefined category, while information extraction (IE) involves identifying and extracting relevant information. In the field of image classification, Convolutional Neural Networks (CNNs) had a predominant role. CNNs are particularly effective in handling complex visual data. Other techniques like SVM's, Logistic regression, random forest etc need the data to be pre-processed which is one of their major drawback; on the other hand CNN has advantage in this regard. CNNs' process is mainly inspired by the human brain, capturing intricate features from images. We aim to provide a detail study on CNNs, and their applications in solving complex image classification tasks.

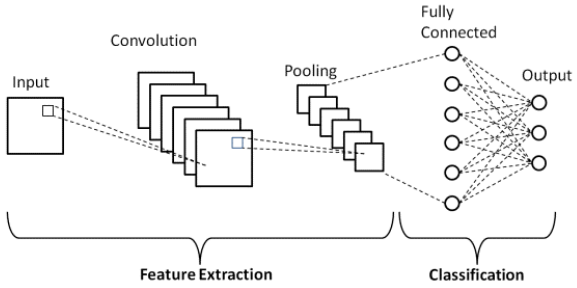

Fig. 1. A basic CNN model

While image classification being a well known domain, data retrieval from images, particularly textual information has been a critical task. One of the important areas is Optical Character Recognition, a technology in Natural Language Processing (NLP), along itwith the capabilities of Long Short-Term Memory (LSTM) networks which can recognize text in images and convert it into a machine-readable format. We can see in the below figure that all the three formats image, PDF or handwritten document can be converted to text. This majorly involves three steps

- *Pre-processing* (Identify content formatting such as columns and tables, font, alignment etc.)
- *Text or character recognition* (Pattern recognition, feature detection, or a combination of both.)
- *Post Processing*

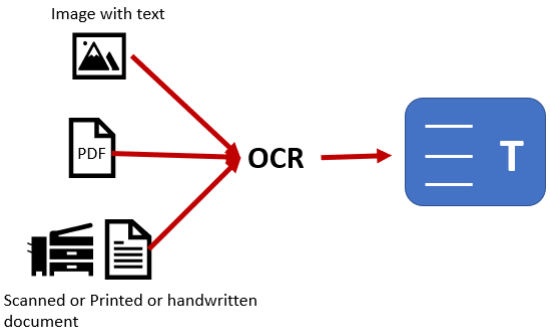

Fig. 2. Optical Character Recognition

We aim to showcase how these technologies OCR and CNNs

can be integrated to get a powerful tool in automated document classification and information extraction.

II. Problem Statement

In modern digital operations, the need for efficient, automated systems for document classification and information extraction has become increasingly important. The current techniques used for verification and data masking involves the tedious process of scanning entire PDF documents, aiming to identify and extract personal authentication information . This process is supported by an Application Programming Interface (API) that applies masking techniques to conceal sensitive information to ensure privacy.

Without knowing the document type, scanning the entire document to extract the required information demands significantly more computation. Moreover it raises privacy and compliance concerns, as sensitive data is processed unnecessarily. Reliance on third-party data extraction services increases both costs and processing delays, as each document sent for external processing incurs fees and lengthens timelines. To address these challenges, we propose a solution that incorporates deep learning techniques for document classification and processing. This approach aims to significantly enhance system efficiency, accuracy, and overall performance. By leveraging deep learning algorithms, the system can automatically identify document types and extract relevant data more effectively, reducing the need for exhaustive document scanning while improving the speed and precision of information retrieval.

We start with preliminary scanning of PDF for Classification and Information Extraction. Through CNN document type is classified, this ensures only classified identification documents go through the process of masking API. By integrating the Tesseract OCR engine, classified document images are automatically processed for data extraction, allowing for faster and more efficient form filling and data aggregation. This leads to significant improvements in overall efficiency and data management. Additionally, this integration enhances data privacy, as sensitive information remains within the organization's control throughout the extraction process.

To conclude the proposed solution for overcoming the identified drawbacks, it is evident that integrating deep learning techniques, such as CNN-based classifier for document classification and Tesseract OCR with LSTM for targeted data extraction, can significantly enhance the current verification and masking process. This approach not only addresses the inefficiencies of exhaustive document scanning but also improves the system's accuracy and operational efficiency by automating the identification and retrieval of sensitive information. Ultimately, this framework offers a scalable and more secure method for document verification and data management in digital operations.

III. Methodology

The process firstly includes classification of collection of different document files having variety of formats through image processing using CNN. This categorization continues furthur processing and gave a path for extraction. Let us first see about the Image/ Document Classification.

A. Image/ Document Classification

Classification of Documents is an essential task for authentication in many of the areas.This involves

1) *Data Pre-processing*
After reading the methodology[1]. Their goal is to convert the type of input (PDF/ image etc) to the one used for further analysis and classification.
   - They had initially used **'fitz' library** for sequential iteration through pages one by one. The content of each page is processed and converted into pixmap representation
   - Then the conversion of the **pixmap** into a standard image format using **'PIL'(Python Imaging Library)** module. This format can be easily applied in various image processing techniques and lays the ground frame for further analysis followed by image resizing to maintain uniformity.
   - The conversion from image to array representation is done using '**image.img_to_array**()' function. This format is the one used for mathematical part.
   - **Pixel Normalization step**: To get all values in the range [0,1] for model convergence as well as stability of training process, it is dividied by 255.
   - More over some of the documents may be blurred, hence sharpening the images before classification can refine the extraction.

   This approach supports us in acheiving the optimized efficiency and accuracy needed and presicion thoughout.

2) *Convolutional Neural Network Model*
CNNs are deep learning models specific to image processing tasks. They extract features directly from images unlike other methods. CNNs use convolutional layers for localized feature extraction followed by pooling layers to down sample features and reduce complexity. Their process include
   - Rigorous iterative fine tuning through optimization, hyperparameters created maximises the accuracy by integrating the data processed before.
   - Preventing Overfitting: when varying filters and layer numbers the model becomes complex starts to memorize data rather than learning. Regularization techniques like(**L1 or L2regularization**) are used.
   - selected optimizers guide weight updates during training for optimal convergence.
   - Firstly the **foundational layer** is **'Conv2D'** layer having 16 filters(3*3 size) activated using ReLU.This layer uses L2 Regularizer for generalisation and next batch normalization layer enhances stability and MaxPooling 2D layer downsamples feature maps to reduce complexity.
   - Same patterns continues for other two layers

- Flatten layer transforms the multidimensional feature maps into a 1D vector, can be taken into full connected two dense layers having some specifications.
- Considering SGD Optimizer with some eta and cross entropy loss iterating through multiple epochs.

With this refined process we go into next phase i.e. Data/Information extraction.

B. Information Extraction

Information Extraction (IE) is the process of extracting useful data from the existing data by using Natural Language Processing (NLP) .It involves the extraction of meaningful and relevent information needed. This process extends beyond classification, focusing on the extracting of specific details, such as text, numbers, or patterns, from various formats like PDF's, images etc.Considering its crucial role, thereby enhancing decision-making, efficiency, and accuracy.These are the steps followed,

1) Tesseract OCR : Tesseract OCR transforms visual data from classified images into actionable text , connects imagebased content and computable information.Tesseract OCR identifies textual patterns regardless of fonts, styles, or orientations, culminating in meaningful data extraction from images.
2) Extraction Process with Pytesseract: Tesseract OCR library, a well-established and widely used tool for OCR tasks. The Python function named 'all_data_extraction' is the important part of this process. It employs Tesseract, along with other libraries such as OpenCV and regular expressions, to meticulously extract critical details.
3) Data Transformation and Parsing: Initially image is transformed and processed using OpenCV to enhance its clarity. Tesseract then converting the processed image into text data. Regular expressions are employed to isolate and extract specific patterns.The function meticulously parses the extracted data, fine-tuning it to eliminate unwanted characters or elements.

In some cases where OCR might fail due to incorrect text orientation, the code includes a mechanism to address this issue. It applies a 90-degree counterclockwise rotation to the image using the 'cv2.rotate' function. Extracted data undergoes a data cleaning process.

By combining OCR, regular expressions, and image rotation, the code ensures reliable extraction even in situations where text orientation, layout, and formatting may vary. In essence, the integration of these functions offers a robust and efficient way for automated data extraction .

REFERENCES

[1] Soham Patel, Dhyey Sanghavi, Prof. Archana Nanade "Modernizing Data Processing: CNNs and OCR for Automated Document Classification and Data Extraction" in 2023 Global Conference on Information Technologies and Communications (GCITC) Karnataka, India. Dec 1-3, 2023

[2] Richa Singh, Vikrant Sharma , Rekha Kashyap "Automated Multi-Page Document Classification and Information Extraction for Insurance Applications using Deep Learning Techniques" in 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. Mar 14-15, 2024

[3] Sammed S. Admuthe, Hemlata P. Channe, "Document Image Classification using Visual and Textual Features" in International Journal of Engineering Research and Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV10IS090183 (This work is licensed under a Creative Commons Attribution 4.0 International License.) 10 Issue 09, September-2021

[4] https://www.egnyte.com/guides/governance/optical-character-recognition

[5] https://www.upgrad.com/blog/basic-cnn-architecture/