POST 33
DATA SCIENCE FOR ALL

Streamlining Bank Statement Data Extraction with Transfer Learning-Enhanced OCR

- **Optical Character Recognition (OCR)** has become an essential tool for **automating document processing in the financial sector.**
- Banks and financial institutions deal with vast amounts of scanned documents, such as bank statements, invoices, and receipts. Extracting relevant information from these documents is critical for streamlining processes like loan approvals, financial audits, and compliance reporting.
- **Transfer learning, a powerful technique in machine learning, allows us to leverage pre-trained OCR models like Tesseract and EasyOCR to enhance the accuracy of bank statement digitization.** This article explores how transfer learning can be applied to OCR for extracting key financial details from scanned bank statements

DineshLal

The Challenge of Bank Statement OCR

Bank statements present unique challenges for OCR systems:

- **Varied Formats:** Different banks have different statement layouts.
- **Noisy Backgrounds:** Watermarks, logos, and security patterns can interfere with text extraction.
- **Multi-Language and Font Variability:** Statements may contain multiple languages, numerical formats, and non-standard fonts.
- **Tabular and Structured Data:** Extracting financial transactions accurately from tables requires more than simple text recognition.

Leveraging Transfer Learning in OCR

Transfer learning enables the adaptation of pre-trained OCR models to domain-specific tasks with minimal additional training. Instead of training an OCR model from scratch, we can fine-tune existing models like Tesseract or EasyOCR for improved performance on bank statements.

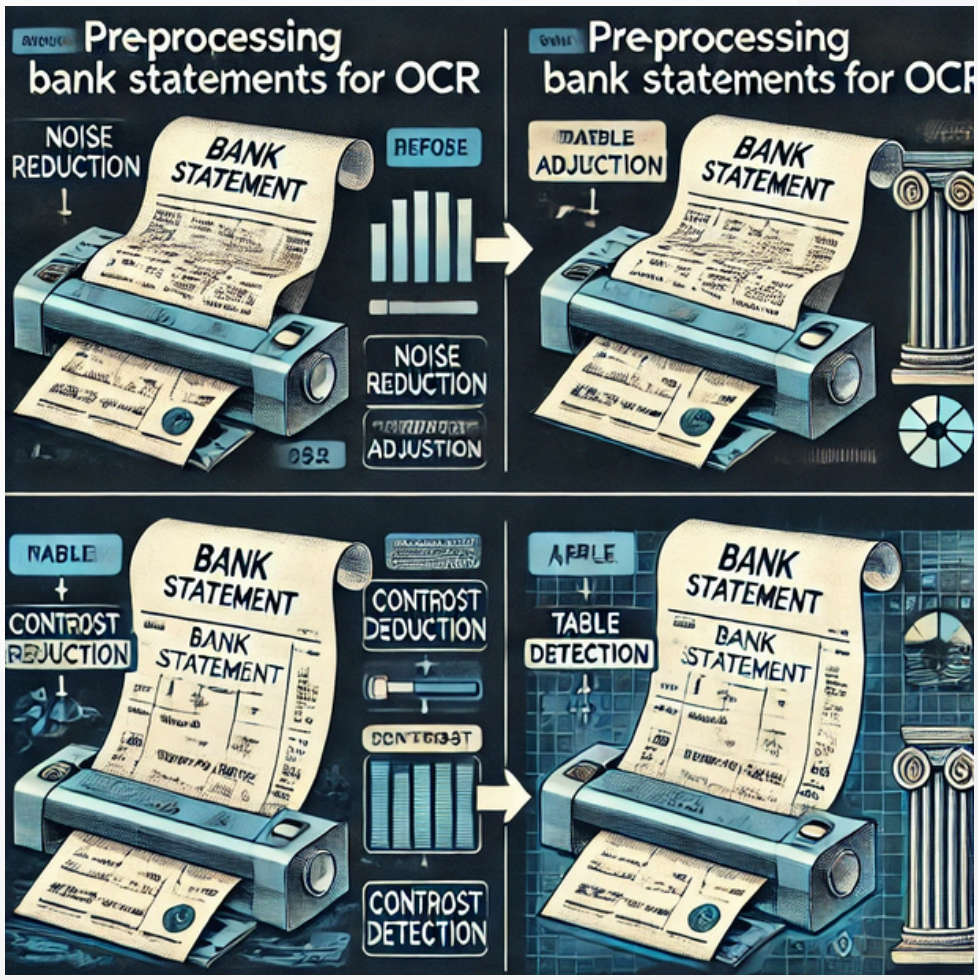
1. Selecting a Pre-Trained OCR Model

- Two widely used OCR models are:
- **Tesseract OCR:** Open-source, rule-based OCR with strong support for multiple languages and layout analysis.
- **EasyOCR:** Deep-learning-based OCR that performs well with noisy and complex document images.

Leveraging Transfer Learning in OCR

2. Preprocessing Bank Statements

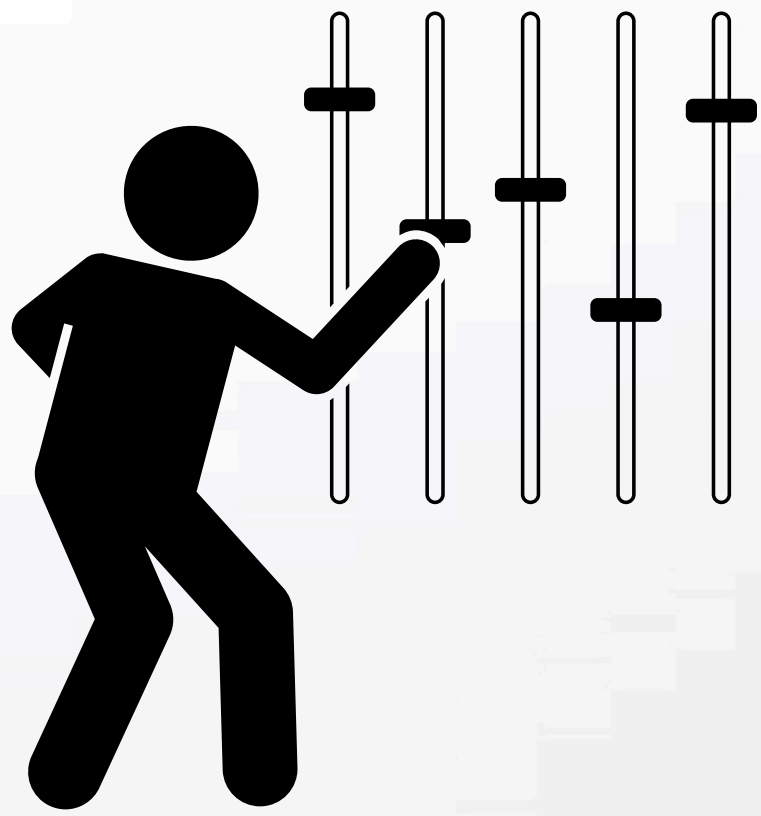
- To improve OCR accuracy, preprocessing techniques can be applied:
- **Noise Reduction:** Removing backgrounds, watermarks, and artifacts using OpenCV.
- **Contrast Enhancement:** Improving text visibility with adaptive thresholding.
- **Table Detection:** Using deep learning models like TableNet or Detectron2 to identify and extract structured tables.



Leveraging Transfer Learning in OCR

3. Fine-Tuning the OCR Model

- Fine-tuning involves training the OCR model with domain-specific data:
- **Dataset Preparation:** Collecting a dataset of scanned bank statements with ground truth annotations.
- **Training Custom Recognition Models:** Using tools like Tesseract's LSTM training or EasyOCR's custom dataset integration.
- **Post-Processing with NLP:** Applying Named Entity Recognition (NER) and regular expressions to extract relevant financial entities like transaction amounts, dates, and account numbers.



Leveraging Transfer Learning in OCR

4. Validation and Optimization

- After training, the model needs to be validated:
- **Accuracy Metrics:** Measuring Word Error Rate (WER) and Character Error Rate (CER).
- **Error Correction:** Implementing domain-specific lexicons and spell correction models.
- **Integration with Financial Systems:** Ensuring seamless data extraction and integration into financial applications.

Example: Extracting Financial Data from a Bank Statement

To illustrate the power of transfer learning in OCR, let's consider a real-world example:

Scenario:

A financial institution wants to automate the extraction of transaction details from scanned bank statements. The goal is to retrieve relevant details such as date, transaction description, amount, and balance.

Implementation Steps:

Preprocessing:

- Convert the scanned document to grayscale.
- Apply adaptive thresholding to enhance text visibility.
- Use OpenCV to detect and remove unnecessary artifacts.

Example: Extracting Financial Data from a Bank Statement

Applying OCR:

- Use EasyOCR to extract raw text from the processed image.
- Identify structured data elements like tables using deep-learning models.

Post-Processing:

- Use Named Entity Recognition (NER) to classify extracted text into relevant fields (date, amount, description).
- Apply regular expressions to standardize transaction formats.
- Cross-validate extracted amounts with total balance for accuracy.

Python Implementation:

With transfer learning, the financial institution successfully achieves 95% accuracy in transaction extraction, reducing manual effort by 80%.


```
IMPORT CV2
IMPORT EASYOCR
IMPORT RE

# LOAD AND PREPROCESS THE IMAGE
DEF PREPROCESS_IMAGE(IMAGE_PATH):
 IMAGE = CV2.IMREAD(IMAGE_PATH, CV2.IMREAD_GRAYSCALE)
 _, PROCESSED_IMAGE = CV2.THRESHOLD(IMAGE, 150, 255, CV2.THRESH_BINARY)
 RETURN PROCESSED_IMAGE

# APPLY OCR
DEF EXTRACT_TEXT(IMAGE_PATH):
 READER = EASYOCR.READER(['EN'])
 TEXT = READER.READTEXT(IMAGE_PATH, DETAIL=0)
 RETURN "\N".JOIN(TEXT)
```

Python Implementation:

With transfer learning, the financial institution successfully achieves 95% accuracy in transaction extraction, reducing manual effort by 80%.


```
# EXTRACT TRANSACTION DETAILS
DEF EXTRACT_TRANSACTIONS(TEXT):
 PATTERN = R"(\D{2}/\D{2}/\D{4})\S+([A-ZA-Z0-9 ]+)\S+(-?\$\D+\.\D{2})"
 TRANSACTIONS = RE.FINDALL(PATTERN, TEXT)
 RETURN TRANSACTIONS

# EXAMPLE USAGE
IMAGE_PATH = "BANK_STATEMENT.PNG"
PROCESSED_IMAGE = PREPROCESS_IMAGE(IMAGE_PATH)
EXTRACTED_TEXT = EXTRACT_TEXT(IMAGE_PATH)
TRANSACTIONS = EXTRACT_TRANSACTIONS(EXTRACTED_TEXT)

FOR TRANSACTION IN TRANSACTIONS:
 PRINT(F"DATE: {TRANSACTION[0]}, DESCRIPTION: {TRANSACTION[1]}, AMOUNT:
{TRANSACTION[2]}")
```

Real-World Applications

Applying transfer learning in OCR for bank statements has several practical benefits:

- **Automated Loan Processing:** Faster verification of financial documents.
- **Fraud Detection:** Identifying inconsistencies in financial transactions.
- **Regulatory Compliance:** Ensuring accurate record-keeping for audits and tax filings.

Summary

- Transfer learning enables financial institutions to significantly improve OCR accuracy for bank statement processing.
- By leveraging pre-trained models like Tesseract and EasyOCR and fine-tuning them with domain-specific data, banks can automate document extraction with higher precision, reducing manual effort and improving efficiency.
- As deep learning continues to evolve, the future of OCR in banking looks promising, with enhanced accuracy and automation capabilities.

Special Thanks to ChatGPT
and Gemini for Content support