Final Report
AI-Powered SMART OCR Scanner and Translation Application

Simon Chaisouang and Venkata Kumari Garlapati
University of Arkansas
shchaiso@uark.edu vg021@uark.edu

1 Introduction

Multilingual translation and optical character recognition (OCR) technology are now essential for reading and scanning text from handwritten notes, physical documents, and public signs. Although cloud-based services are frequently used in current solutions, they have drawbacks such as reliance on the Internet, sluggish processing speeds, and possible privacy issues. There is a growing need for novel solutions that run directly on edge devices due to the growing demand for systems that are faster, more secure, and offline capable. To accomplish end-to-end handwritten text detection and language translation without relying on cloud infrastructure, this research suggests an AI-powered pipeline. The system processes input data in real time using cutting-edge transformer models, such as mBART-50 for effective multilingual translation and TrOCR for precise handwritten text recognition.The pipeline combines translation, recognition, and sophisticated preparation methods into a unified framework that is tailored for edge devices. This method guarantees that users can use the system without an active internet connection while also improving accessibility and privacy.

Applications in sectors such as logistics, healthcare, education and accessibility are made possible by developing lightweight, edge-optimized AI models, which close the gap between sophisticated machine learning capabilities and resource-constrained contexts. To satisfy the increasing need for AI-powered automation in practical settings, this project establishes the foundation for high-performance, scalable, privacy-preserving OCR and translation systems.

2 Problem Statement

The growing demand for low-latency, real-time text translation and recognition has sparked interest in edge-based AI solutions [3]. Cloud-based solutions are the foundation of traditional OCR and translation services, raising issues with response speed, network dependence, and data protection. To lessen dependency

on cloud services and preserve high accuracy and speed, this project intends to create an AI-powered OCR scanner with translation capabilities that can function well on edge devices.

The cloud APIs used by most OCR and translation systems today are ineffective for real-time use, jeopardize user privacy, and need consistent internet access. A reliable on-device solution that functions well on low-power devices and guarantees speed, privacy, and offline capability is required.

3 Literature Review

The challenge of text recognition has been around for a while, especially in the context of document digitization [3]. Conventional OCR techniques use recurrent neural networks (RNNs) to generate text at the character level and convolutional neural networks (CNNs) to interpret images. To increase accuracy, language models are also frequently employed as post-processing procedures.

Recent works, such as the Show, Attend, and Read model, introduced attention-based methods that significantly improve OCR accuracy on irregular text by dynamically focusing on different regions of an image. [2]

Existing research in multilingual translation and optical character recognition (OCR) has advanced significantly by utilizing both conventional and contemporary methods. Traditional OCR systems, like Tesseract, rely on optical pattern matching and rule-based algorithms, but because of their restricted flexibility, they have trouble processing intricate handwriting and multilingual text. Cloud-based OCR and translation services, such as Microsoft Azure OCR and Google Cloud Vision, use strong machine learning models to achieve high accuracy, but also require internet connectivity, which presents privacy issues.

With models like TrOCR excelling at handwritten and printed text recognition and mBART-50 providing multilingual translation across more than 50 languages, recent advances in transformer-based models have completely changed these tasks. Furthermore, MobileNet and TensorFlow Lite are examples of edge-based AI models that seek to provide effective and portable solutions for on-device applications. The incapacity of conventional OCR techniques to handle a variety of handwriting and languages, privacy concerns with cloud-based systems, and computational difficulties for on-device models are some of the drawbacks of current solutions. By incorporating cutting-edge transformer models into an edge-based pipeline, our project fills these gaps and offers an accurate, effective, and cloud-independent solution.

4 Proposed Solutions

To overcome reliance on cloud-based solutions, this project presents an end-to-end pipeline for handwritten text recognition and language translation that is powered by AI and runs solely on edge devices. By utilizing cutting-edge lightweight transformer models, the suggested solution overcomes the drawbacks

of current approaches, including internet dependence, latency, and privacy issues.

- The proposed solution's salient features include:
  - **Using the device to deploy:**
    TrOCR and mBART-50 are optimized for edge platforms, which guarantees effective operation on devices with limited resources, including embedded systems or mobile phones.
  - **Offline Functionality:**
    The solution can function in locations without internet connectivity due to its lack of cloud dependencies, which improves security and usability.
  - **Real-Time Performance:**
    Applications like document digitalization, language translation, and accessibility tools can benefit from the solution's ability to process input in real-time and produce results instantly.
  - **Scalability:**
    The pipeline's modular architecture enables the addition of other languages or functionalities, such domain-specific translations or real-time handwriting style adaption.

The suggested approach establishes the groundwork for a safe, quick, and scalable OCR and translation system designed for edge deployment by fusing the advantages of mBART-50 for flexible multilingual translation with TrOCR for precise handwriting recognition.

4.1 Transformer-based OCR, or TrOCR

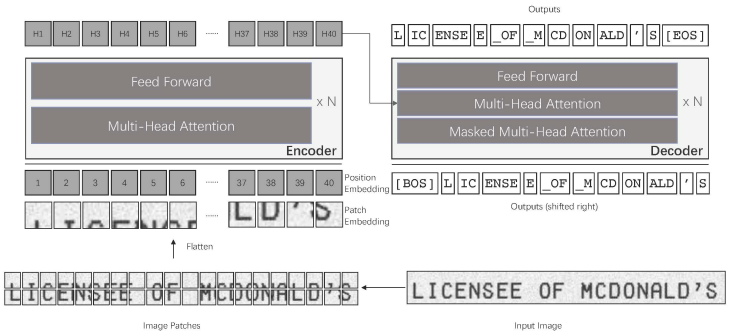

Figure 1: Pipeline of TrOCR Model

A state-of-the-art transformer model for reliable optical character identification. Ability to use its attention mechanism to effectively handle variances in

handwriting styles and layouts, as well as identify and recognize handwritten text with accuracy. The pipeline for the TrOCR model can be broken down in the following:

- **Input Stage:**
  - **Input Image:** A text image is used as the system's starting point. The text `"LICENSEE OF MCDONALD'S"` is an example of input.
  - **Patch Extraction:** The input image is preprocessed by splitting it into equal-sized, non-overlapping patches. Each patch may represent letters or parts of letters. This step is crucial as Transformers process sequences rather than raw 2D images.
  - **Patch Embedding:** Each patch is flattened into a 1D vector and mapped into a high-dimensional feature space via a learned linear embedding. Positional embeddings are added to retain spatial relationships between patches.
- **Encoder:**
  - **Feed-Forward Layers:** Linear layers convert patch embeddings into abstract representations.
  - **Multi-Head Attention:** This mechanism allows the model to relate different patches, capturing global patterns—essential when characters span multiple patches.
  - **Output:** The encoder produces a sequence of high-dimensional feature vectors (H1, H2, ..., H40) encoding the semantic and structural information of the image.
- **Decoder:**
  - **Decoder Input:** A sequence of tokens initializes the decoder, beginning with a special `[BOS]` token. The model predicts the next token using both the encoded image features and previously generated tokens.
  - **Masked Multi-Head Attention:** Ensures the decoder only attends to earlier tokens, which is critical for autoregressive decoding.
  - **Decoding Text:** The decoder uses positional context and encoder outputs to generate text token by token, e.g., producing `"L"`, then `"I"`, `"C"`, and so on, until the `[EOS]` token.
- **Output:**
  - The final result is a complete transcription of the input image, such as `"LICENSEE OF MCDONALD'S"`. The autoregressive process enables robust text generation even from noisy inputs.
- **Key Features of the Transformer OCR Model:**

- **Patch-Based Processing:** Allows flexible handling of text images by converting them into sequences.
- **Encoder-Decoder Architecture:** The encoder extracts visual features; the decoder maps them to readable text.
- **Positional Embeddings:** These provide ordering context to the otherwise position-agnostic Transformer.
- **Attention Mechanisms:** Encoder's attention identifies global visual dependencies; the decoder's masked attention ensures proper token generation.
- **Autoregressive Decoding:** Text is generated step-by-step using prior outputs, ensuring coherence and context preservation.

4.2 Multilingual Translation, or mBART-50

A multilingual transformer model that has been optimized for accurate translation in more than 50 languages [5]. It allows for the accurate translation of detected text into several languages while preserving contextual information.

5 Experiments

5.1 Datasets

The datasets used for this study consist of the IAM Handwriting Dataset [4] and the WMT19 Dataset [1]. The IAM dataset was used for training the TrOCR model for text recognition. The dataset consists of 1,539 scanned pages of handwritten English text which comprises of 13,353 lines of text and 115,320 words total. The handwritten samples were provided by 657 different writers, offering a diverse range of handwriting styles. Each sample is in the form of a PNG file and are greyscale.

The WMT19 dataset orginates from the 2012 Conference on Machine Translation, an event where researchers come together to evaluate and advance machine translation systems. The consist of 18 different language pairs in the form of parallel text which are sentences that are aligned between two different languages. For this study, we focus on the English to Chinese and English to French translations for training our mBART-50 model.

5.2 Data Preprocessing

Each dataset was split into 80/20 for training and testing of each model. For the IAM datase, a file key in the format of a text file that contains a list of all the image titles pertaining to a given subset was created in order for the source code to read in each of the raw png image files to be used in our recreation of the pipeline. Since the mBART-50 model cannot process raw stings, the WMT19 dataset was tokenized to convert the raw sentence pairs into sequences of token IDs to map each input ID to a corresponding label.

5.3 Experiment Results

Figure 2: Sample Results of OCR Scanner Application

The image provided above provides some sample results from our AI-Powered OCR Scanner and Translation application. Our model performed really well when it came to text recognition. For the translation portion of our application, it was able to provide sufficient translation between the English language and the target language specified in the applications with some noticeable errors.

For the French translation in the top-right sentence, it recognized the work "please" as "merci" which means thank you in French. The proper translation for "please" in French would be "s'il te plaît". In the bottom-right sentence, the portion "The first day of the week." was translated as Monday in Chinese. This mistranslation is likely due to the direct transaltion behind the word "Monday" in Chinese being "day one of the week". These translation errors are a representation of one of the shortcomings of our translation model, how it recognizes patterns rather than the underlying meaning of each sentence.

6 Conclusions

From this study, we were able to provided a foundation for AI-Powered OCR Scanner and Translation capable of real-time and offline results. Our application was able to successfully display the integration of TrOCR for accurate handwritten text recognition and mBART-50 for Chinese translations of recognized English text to form a full OCR-to-translation pipeline.

For future work in this study, we could also consider extending the pipeline to account for text detection to be able to capture and process text from various sources rather than solely from a single line of text. To address the shortcomings of the minor translation errors in our model, we could also consider fine-tuning the translation model (mBART-50) with additional datasets to process for accurate translations of English text. This application can also be extended to the Edge for mobile use such as iOS or Andriod. These considerations would increase the relevancy of our app to be useful in real-world scenarios.

References

[1] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, A. Névéol, M. Neves, M. Post, M. Turchi, and K. Verspoor. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, 2019. Association for Computational Linguistics.

[2] H. Li, P. Wang, C. Shen, and G. Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. *arXiv preprint arXiv:1811.00751*, 2018.

[3] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*, 2021.

[4] U.-V. Marti and H. Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[5] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.