MSdocTr-Lite: A Lite Transformation for Full Page Multi-script Handwritting Recognition

Kim Long Hoang   
May 31, 2024

***Abstract*— The Transformer architecture excels in pattern recognition but requires large datasets for training and validating. In Handwritten Text Recognition (HTR), gathering extensive labeled data is challenging and costly. This paper presents a lightweight transformer for full-page multiscript handwriting recognition, addressing data scarcity by enabling training on reasonably sized datasets with- out external data. It learns page-level reading order using a curriculum learning strategy, eliminating line segmentation errors and reducing segmentation annotation needs. Additionally, it supports easy adaptation to other scripts through transfer learning with page-level labeled images. Experiments on various datasets (French, English, Spanish...) demonstrate the model’s effective- ness. Keywords: Seq2Seq model, page-level recognition, Handwritten Text Recognition, Multi-script, Transformer, Transfer Learning**

# Introduction

Handwritten Text Recognition (HTR) converts scanned handwritten documents into machine-readable text but faces challenges due to handwriting variability and segmentation issues. Traditional methods struggle with segmentation, leading to errors. Recent approaches using deep learning, like trans- former models, aim to recognize text at the page level, avoiding segmentation. This paper proposes a lightweight transformer model trained with a curriculum learning strategy, which is efficient, adaptable to various scripts, and performs well across multiple languages. The structure includes related work, proposed approach, experimental results, and conclusions.

# Theoretical Background

## Background on Handwritten Text Recognition (HTR)

In the past, Various approaches have been developed but it stil have a lot of challenging problem due to the high variability in writing style, illegible handwriting, and the degradation of document quality over time. Early methods focused on character-level segmentation, but they struggled with the inherent difficulties of cursive handwriting. To address this, word-based and line-based segmentation techniques were introduced, each offering improvements yet still facing.

significant hurdles.

## *Overview of current HTR approaches*

* Character-Level Segmentation

A close-up of a paper

Description automatically generated

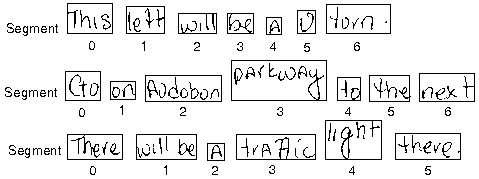
Figure 1:

What it does: Extracts individual character images from documents for recognition.

Weaknesses: Performance is generally low due to the difficulty in segmenting cursive and unconstrained handwriting accurately.

Proposed Solution: Move towards word-based or line-based segmentation to reduce the segmentation errors at the character level.

* Word-Based Segmentation



What it does:

Extracts word images from documents before feeding them into the HTR system.

Weaknesses:

Faces issues due to irregular inter-word and intra-word spaces, making segmentation inconsistent

Proposed Solution:

Adopt text line recognition to utilize more context and reduce the inconsistencies faced in word segmentation.

* Line Segmentation

What it does:

Segments documents into individual lines to surpass inconsistencies between words and provides more context for recognition.

Weaknesses:

Challenges arise from non-uniform text line skew/slant or closely situated and touching text lines, affecting recognition performance.

Proposed Solution:

Explore paragraph or page-level recognition to avoid intermediate segmentation steps and leverage deep learning models.

* Paragraph-Level Recognition with MDLSTM and Attention

What it does:

Uses Multi-Dimensional Long Short-Term Memory Recurrent Neural Networks (MDLSTM) and attention mechanisms for paragraph-level HTR.

Weaknesses:

High memory requirements, lack of GPU acceleration for training, and intractable inference time led to abandonment.

Proposed Solution:

Investigate transformer-based architectures and curriculum learning strategies to improve efficiency and scalability.

* Sequence-to-Sequence (Seq2Seq) Models Based on Transformer Architecture

What it does:

Utilizes the transformer architecture for page-level handwritten document text recognition.

Weaknesses:

Requires a massive amount of annotated data for training.

Long feature sequences lead to extensive training times and high GPU resource consumption.Difficult to port to low-resource languages due to the need for extensive data for retraining.

Proposed Solution:

Develop a lite transformer model that requires fewer parameters and can be trained on standard GPUs. Implement curriculum learning to efficiently train the model with limited annotated data and facilitate transfer learning to adapt to different scripts.

## Motivation for developing a lite transformer model

Motivated by the need for efficient and effective handwritten document recognition, we propose a lite transformer model for page-level handwritten text recognition. This model uses a limited number of parameters and can be trained without external data. Employing a curriculum learning strategy, the model learns reading order and scales to large text images. This strategy is applied once, making the model adaptable to different scripts with minimal additional training. This architecture requires less memory, allowing training on standard GPUs. Key contributions include:

**Key Contributions:**

* An end-to-end lite transformer-based model for page-level handwritten text recognition, avoiding early segmentation errors and leveraging larger context.
* Curriculum learning strategy to reduce the need for extensive annotated data and enable the model to learn reading order at the page level.
* Adaptability to different scripts using a simple transfer-learning process with page-level labeled images.
* Extensive comparative experiments demonstrate effectiveness across multiple scripts and languages (e.g., English, French, Spanish, Arabic).

By addressing the weaknesses of previous approaches, the lite transformer model aims to offer a more efficient and adaptable solution for handwritten text recognition.

# Related Work

## Review of line level HTR systems

The Line-level HTR system are specialized tools designed to recognize and transcribe text at a line level, meaning they process one line of text at a time. These systems are crucial for digitizing handwritten documents, enabling easier search,

editing, and storage.

The Key Components of line-Level HTR-systems are:

* Preprocessing
* Feature Extraction
* Recognition
* Postprocessing  
  The problem of mapping transcript to images has motivated some research in the past decade, either for the alignment of Optical Character Recognition (OCR) output with book content (e.g. in [5], [6]), or for mapping the transcription of historical documents to segmented words or lines (such as [7], [8]).

## Review of Page level

Archives and libraries around the globe possess billions of historical manuscripts. Although numerous documents have been digitized as image, access is still restricted since their contents aren’t available in a symbolic format that enable modern textual process like editing, indexing and retrieval. Handwritten Text Recognition (HTR) plays a pivotal role here, seeking to offer automatic transcription methods for these documents [11].

While line-level HTR systems focus on individual lines of text, page-level HTR systems offer serval advantages by considering the entire page. These advantages include better contextual understanding, efficiency in processing, handling complex layouts, scalability for large projects, improved accuracy through error correction and consistent formatting and the ability to provide integrated, end-to-end solutions. These benefits make page-level HTR system particularly valuable for complex and large-scale digitalization efforts.

The Author in paper [12] has also made an experiment to show example of misleading WER evaluation, caused by wrong reading order due to text-line detection flaws. While all the words in the automatic transcript are perfectly correct, the WER is 70% (13% matching words, 13 substitution errors, 4 insertions and 4 deletion). The image is part of a page of the

Bentham Papers collection  
  
A close-up of a paper

Description automatically generated

Considering the limitations of previous works, we pro-

pose in this paper MSdocTr-Lite, an end-to-end lite

transformer-based model for multi-script handwritten doc-

ument recognition

# Proposed Approach

# Experimental Results

## Experimental Setup

* Database:

To evaluate the quality, strengths, and weaknesses of this system, we will use databases where we know the ground truth for both line segmentation and transcription. These databases are publicly available and widely used in automatic text recognition research. The Rimes database [1] includes a training set of 1,500 images of handwritten French paragraphs and an evaluation set of 100 images. The IAM database [2] provides 747 images of handwritten English documents for training, 116 for validation, and 336 for evaluation. Figure 3 shows examples of these images. We conducted experiments on portions of the databases that were not used to train the recognizers.

## Result and analysis

## Ablation Studies

Discussion

* Analysis of the results in the context of data scarcity
* Advantage of the proposed lite transformer model
* Limitations

Conclusions

A diagram of a process

Description automatically generated

Figure **Error! Use the Home tab to apply 0 to the text that you want to appear here.**‑

# References and Footnotes

## A. References

REFERENCES

[1] E. Augustin, M. Carre, E. Grosicki, J.-M. Brodin, E. Geoffrois, and ´ F. Preteux, “RIMES evaluation campaign for handwritten mail processing,” in Proceedings of the Workshop on Frontiers in Handwriting Recognition, no. 1, 2006.

[2] U.-V. Marti and H. Bunke, “The IAM-database: an English sentence database for offline handwriting recognition,” International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 39–46, Nov. 2002.

[5] S. Feng and R. Manmatha, “A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books,” Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries- JCDL ’06, pp. 109–118, 2006.

[6] I. Z. Yalniz and R. Manmatha, “A fast alignment scheme for automatic ocr evaluation of books,” in Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011, pp. 754–758.

[7] J. Rothfeder, R. Manmatha, and T. M. Rath, “Aligning Transcripts to Automatically Segmented Handwritten Manuscripts,” in Proc. 7th Int. Workshop on Document Analysis Systems, 2006, pp. 84–95.

[8] E. Kornfield, R. Manmatha, and J. Allan, “Text alignment with handwritten documents,” in First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings. IEEE, 2004, pp. 195–209.

[11] G. Muehlberger, L. Seaward, M. Terras, S. A. Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Dejean, ´ M. Diem, S. Fiel, et al. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. Journal of documentation, 2019

[12] End-to-End Page-Level Assessment of Handwritten Text Recognition Enrique Vidal,1,2 Alejandro H. Toselli,1 Antonio R´ıos-Vila,3 Jorge Calvo-Zaragoza3 1PRHLT Research Center, Universitat Politecnica de Val ` encia ` 2Valencian Graduate School and Research Network of Artificial Intelligence (valgrAI) 3U.I. for Computer Research, University of Alicante