Philipps-University Marburg
FB12: Department of Mathematics and Computer Science

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 handwritting recognition, addressing data scarcity by enabling training on reasonably sized datasets without 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 effectiveness. Keywords: Seq2Seq model, page-level recognition, Handwritten Text Recognition, Multi-script, Transformer, Transfer Learning

Contents

List of Figures

List of Tables

1 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 transformer 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.

2 Theoretical Background

2.1 Background on Handwritten Text Recognition (HTR)

Handwritten Text Recognition (HTR) converts scanned handwritten documents into machine-readable text. It's challenging due to diverse writing styles, poor document quality, and unique script properties. Early HTR methods relied on character or word segmentation, which struggled with cursive and inconsistent handwriting. More recent approaches focus on text line recognition, achieving better performance but still facing issues with line skew and spacing. Advances in deep learning have led to paragraph or page-level recognition, but these methods require significant computational resources and annotated data. This paper proposes a lightweight transformer model that uses curriculum learning and standard GPUs, making it efficient and adaptable to various scripts.

2.2 Overview of current HTR approaches

Current Handwritten Text Recognition (HTR) approaches include:

- Character-Level Segmentation: Struggles with cursive writing accuracy.
- Word-Level Segmentation: Faces issues with irregular spacing between words.
- Text Line Recognition: State-of-the-art performance but challenges with skew/slant lines.
- Deep Learning-Based Methods: Use MDLSTM and transformers, requiring extensive data and computational resources.

The proposed solution is a lightweight transformer model for page-level HTR, trained with curriculum learning, needing fewer resources, and adaptable to various scripts.

2.3 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. Our architecture requires less memory, allowing training on standard GPUs. Key contributions include:

- An end-to-end lite transformer model avoiding early segmentation errors.
- Curriculum learning strategy for efficient training with limited annotated data.
- Adaptability to other scripts using simple transfer learning.
- Validation across multiple scripts and languages, confirming the model's effectiveness.

3 Related Work

3.1 Review of line-level HTR systems

Early character and word segmentation methods. Line-level recognition and its advantages over word-based approaches.

3.2 Review of page-level HTR systems

Encoder-decoder architectures with attention mechanisms. Limitations of existing transformer-based models in terms of data requirements and computational resources.

4 Proposed Approach

4.1 Model Architechture

Description of the lite transformer model.

- Transformer-Encoder: Feature extraction and representation.
- Transformer-Decoder: Sequence-to-sequence mapping.

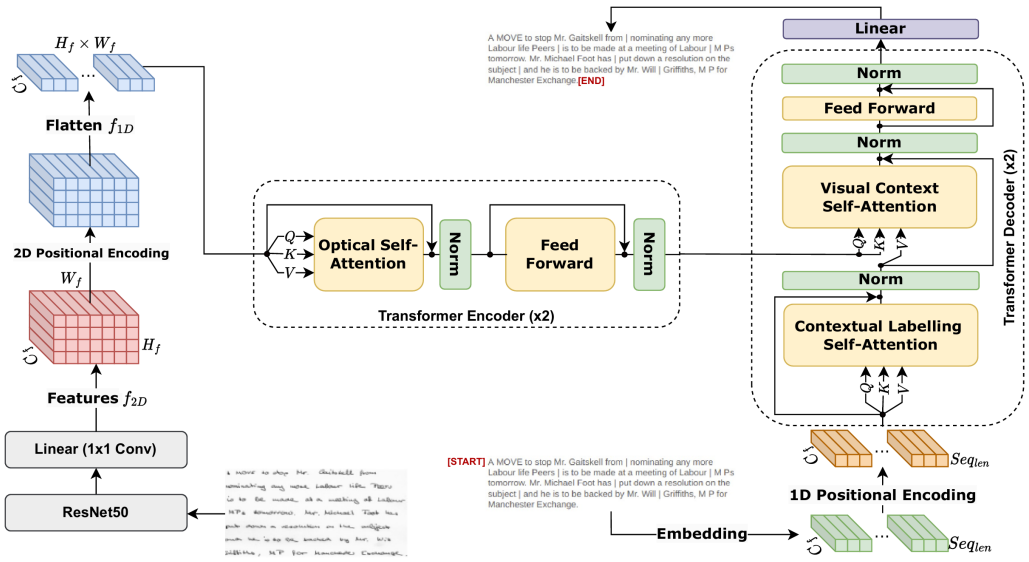

Figure 1: Overview of the proposed architecture. The lite transformer is composed of a transformer encoder combining convolutional layers and transformer layers and of a transformer decoder.

4.2 Curriculum Learning Strategy

Importance of curriculum learning in training transformers with limited data. Three-stage training process: Initial training with small blocks of text. Fine-tuning with larger, more complex blocks. Final tuning with full page-level data.

4.3 Transfer Learning

Adapting the model to different scripts using transfer learning. Process and benefits of transfer learning in multi-script HTR.

5 Experimental Results

5.1 Experimental Setup

- Description of datasets used (e.g., IAM, RIMES, Esposalles, KHATT).
- Evaluation metrics and baseline comparisons.

5.2 Results and Analysis

- Performance of the lite transformer on different datasets.
- Comparison with state-of-the-art models.
- Impact of curriculum learning and transfer learning on model performance.

5.3 Ablation Studies

- Impact of different components of the transformer.
- Detailed analysis of fine-tuning stages and their contributions.

6 Discussion

- Analysis of the results in the context of data scarcity and model efficiency.
- Advantages of the proposed lite transformer model.
- Lower computational requirements.
- Improved adaptability to different scripts.
- Limitations and potential areas for improvement.

7 Conclusion

- Summary of key findings and contributions.
- Implications for future research in HTR and deep learning.
- Final thoughts on the impact of the proposed model on multimedia data analysis

8 References

Comprehensive list of all sources cited in the paper, including seminal works and recent studies in HTR and deep learning.

9 Appendices (if applicable)

- Additional data, code snippets, or detailed experimental results.
- Supplementary material supporting the main content of the paper.

References