Towards Efficient End-to-End Multi-Scale Handwritten Document Recognition via Deep Learning

Abstract - Handwritten Text Recognition (HTR) remains a fundamental challenge in document analysis, particularly when processing complex layouts and historical manuscripts. This thesis presents a systematic progression in HTR technology, addressing critical challenges across multiple scales while optimizing the balance between accuracy, efficiency, and practical deployability. Our research advances through four key stages, each building upon previous achievements to handle increasing document complexity.

Our initial contribution introduces an innovative line-level approach, combining a split-attention CNN backbone (ResNeSt101) with transformer-based sequence modeling. This foundation achieves enhanced feature extraction for diverse handwriting styles and improved recognition accuracy for out-of-vocabulary words through character-level processing. Building upon this, we develop a segmentation-free pipeline for paragraph-level recognition, introducing joint attention mechanisms that handle complex layouts without explicit line segmentation and enable end-to-end trainable document understanding.

Advancing to multi-page document analysis, we propose the Hierarchical Attention Document Network (HAND), featuring a Multi-Scale Adaptive Processing framework that dynamically adjusts to document complexity. HAND incorporates advanced convolutional blocks with Gated Depth-wise Separable and Octave Convolutions, coupled with hierarchical attention mechanisms combining memory-augmented and sparse attention for efficient sequence processing. This innovation significantly advances the processing of complex historical manuscripts with varying layouts and structures.

Our final contribution addresses practical deployment challenges through HTR-JAND (Joint Attention Network and Knowledge Distillation), achieving a 48%

parameter reduction while maintaining competitive accuracy. We introduce a novel Combined Attention Layer integrating Multi-Head Self-Attention with Proxima Attention, reducing computational complexity by 35%. Through effective knowledge distillation and curriculum learning strategies, we achieve 51% faster inference and 40% improvement in cross-dataset generalization.

Extensive evaluations on benchmark datasets including RIMES, IAM, Bentham, Saint Gall, Washington, and READ 2016 demonstrate state-of-the-art performance across multiple scales. Our comprehensive approach encompasses transfer learning, synthetic data generation, and domain-adaptive techniques, resulting in robust performance across diverse historical periods and manuscript styles. All implementations and pre-trained models are publicly available to facilitate reproducibility and advance the field of document understanding.

Keywords: *Handwritten Text Recognition, Attention Mechanisms, Convolutional Neural Networks, Transformers, Knowledge Distillation, Historical Document Analysis, End-to-End Recognition, Multi-Scale Document Processing*