Deep learning with multilingual Transformers for image-to-text recognition

Position	Master's intern
Location	IRISA, SHADoc team 263 avenue du Général Leclerc, 35000 Rennes (France)
Duration	6 months
Desired start date	February 1, 2026
Supervision & contacts	Denis Coquenet (denis.coquenet@irisa.fr) Bertrand Couasnon (bertrand.couasnon@irisa.fr) Yann Soullard (yann.soullard@irisa.fr)
Team website	https://www-shadoc.irisa.fr/

Keywords

> Multilingual Handwritten Text Recognition, Deep Learning, Transformers, Image-to-sequence

Subject

Handwritten Text Recognition (HTR) is the process of automatically extracting handwritten text from a digitized image and converting it into machine-readable text. The historical approach to tackle this task, still in the majority today, consists of a multi-stage paradigm [1]. A document image is segmented into text lines, which are then ordered so as to preserve a coherent reading order, then recognized and reassembled to obtain the overall transcription of the document. This multi-step approach comes with significant drawbacks: 1) the segmentation step relies on a neural network that needs to be trained, requiring additional segmentation annotations, 2) the ordering step is generally a hand-crafted algorithm that needs to be adapted to the layout type and therefore to each dataset, 3) the recognition of the individual lines is carried out independently, limiting the context modelling during prediction, 4) errors accumulate in each of these steps. More recently, the Document Attention Network (DAN) [2] was proposed as the first end-to-end approach for HTR at page level, achieving similar recognition performance compared to the multi-step approach. The DAN includes a transformer-based [3] decoder to perform a character-level autoregressive prediction process. Indeed, the transformer's attention mechanism is used to generate an implicit segmentation of characters, one after the other, making the model learn the reading order of the document through a text-only supervision.

Currently, the main limitation of such approach is that they are limited to a single language and document layout: they are trained and evaluated only on one dataset. The goal of this internship is to go a step forward, *i.e.*, to be able to recognize multiple languages and multiple layouts with a single model. To this aim, we will focus on different research axes. 1) Encoding the language information in the target sequence to guide the network and make reading order easier to understand. The goal here is to evaluate the impact of adding a language identification sub-task in the process. 2) Designing a multilingual training strategy. We want to find the best learning strategy for managing multilingual documents: is it better to have incremental learning, one language at a time, and if so, how can we avoid forgetting previously recognized languages? 3) Evaluating the mixture-of-expert [4] strategy for this task, which consists in dedicating a part of the model to each language, as it was successively applied to Automatic Speech Recognition [5] and Large Language Models [6] (DeepSeek, Mixtral).

References

[1] J. Chung and T. Delteil, "A computationally efficient pipeline approach to full page offline handwritten text recognition," in *Second International Workshop on Machine Learning, WML@ICDAR*, pp. 35–40, IEEE, 2019.

[2] D. Coquenet, C. Chatelain, and T. Paquet, "Dan: a segmentation-free document attention network for handwritten document recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5998–6008, 2017.

[4] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts in large language models," *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[5] S. Cao, X. Wang, Y. Zhang, X. Zhang, and L. Ma, "M-moe: Mixture of mixture-of-expert model for ctc-based streaming multilingual asr," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[6] K. T. Chitty-Venkata, S. Howland, G. Azar, D. Soboleva, N. Vassilieva, S. Raskar, M. Emani, and V. Vishwanath, "Moe-inference-bench: Performance evaluation of mixture of expert large language and vision models," *arXiv preprint arXiv:2508.17467*, 2025.

SHADoc team

The Shadoc team (Systems for Hybrid Analysis of DOCuments) focuses on modelling man-made data for written communication: handwriting, gesture (2D and 3D), and documents, under various aspects: analysis, recognition, composition, interpretation.
The objective is to achieve a continuum between paper and digital documents with a certain readability. We mainly focus on the following topics:

> Intelligent recognition of handwritten content: documents, writings, gestures;

> Analysis of the semantic/structural content: document structure, stages of production of diagrams,drawings, musical scores, sketches, architectural plans;

> Design of new AI, combining recognition and analysis: offer enriched experiences for digital humanities or e-education.

The roadmap of the Shadoc team is on the frontier of several research axes: Pattern Recognition, Machine Learning, Artificial Intelligence, Human-Machine Interaction, Uses and Digital Learning.
Our research is characterized by the hybridization of several AI approaches: two-dimensional grammars,deep learning, fuzzy inference systems... This hybridization aims at guaranteeing, beyond performance,important aspects such as: explicability, genericity, adaptability, data frugality.
Beyond hybridization, the originality of this research is to focus on user interaction. This strategy aims at answering the limits of the current approaches which are based on non-interactive treatments. The concept is to reinforce the decision processes by relying on the implicit validations or explicit corrections of a user to avoid the propagation of errors throughout the analysis. The notions of interpretation, adaptation and incremental learning are at the heart of this research, the objective being to design efficient, robust and self-evolving system.

IRISA lab

IRISA is today one of the largest French research laboratory (more than 850 people) in the field of computer science and information technologies. Structured into seven scientific departments, the laboratory is a research center of excellence with scientific priorities such as bioinformatics, systems security, new software architectures, virtual reality, big data analysis and artificial intelligence.
Located in Rennes, Lannion and Vannes, IRISA is at the heart of a rich regional ecosystem for research and innovation and is positioned as the reference in France with an internationally recognized expertise through numerous European contracts and international scientific collaboration.
Focused on the future of computer science and necessarily internationally oriented, IRISA is at the very heart of the digital transition of society and of innovation at the service of cybersecurity, health, environment and ecology, transport, robotics, energy, culture and artificial intelligence.
IRISA is a joint-venture resulting from the collaboration between nine institutions, in alphabetical order: CentraleSupélec, CNRS, ENS Rennes, IMT Atlantique, Inria, INSA Rennes, Inserm, Université Bretagne Sud, Université de Rennes. From this collaboration is born a force that comes from women and men who give the best of themselves for fundamental and applied research, education, exchanges with other disciplines, transfer of know-how and technology, and scientific mediation.