EA 4108
www.litislab.eu

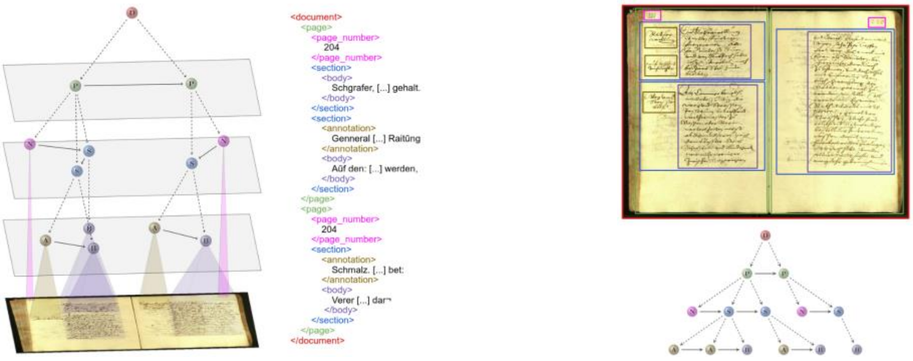
FR CNRS 3638
www.normastic.fr

Master internship position
Synthetic Document Generation for Document Attention Network

For many years, document image analysis and recognition were considered complex tasks decomposed in two independent sub-tasks: layout analysis and text recognition. Layout analysis aims at detecting every image components such as text blocs, tables, images, graphics, signatures… Text recognition is devoted to the recognition of text inside text blocs (paragraphs). Recently, deep architectures have improved the state-of-the-art performance for both sub-tasks. But training each task independently requires twice the amount of annotation: the position (bounding box) of each element in the image; the transcription of text blocs with text line bounding boxes. Moreover, by training the two components independently there is no guaranty that the sequential pipeline is optimum. In any case, the pipeline cumulates the errors of each component. In addition, when moving from one type of document to another, each component requires its specific optimization/adaptation phase.

To overcome these problems and ease the training and adaptation phase of the processing pipeline, we proposed the Document Attention Network that achieves the document processing pipeline end-to-end for the first time (https://arxiv.org/abs/2203.12273). The system, composed of a fully convolutional encoder and a transformer decoder, is trained without any segmentation ground truth. The document structure is encoded with specific TAGS so that the whole document transcription is an Xml-like document with layout TAGS (see figure below). The system is trained to read the document in a chosen reading order that is defined in the ground truth. The complex DAN architecture is pre-trained on synthetic documents first, and then trained on real data following a curriculum learning strategy. The DAN achieves state of the art performance but is very dependent of its pre-training phase, and of the quality of the synthetic training examples generated.

This internship will explore some possible strategies to generate efficient training samples for the DAN.



```
<document>
  <page>
    <page_number>
      204
    </page_number>
    <section>
      <body>
        Schgrafer, [...] gehalt.
      </body>
    </section>
    <section>
      <annotation>
        Genneral [...] Raitũng
      </annotation>
      <body>
        Aũf den: [...] werden,
      </body>
    </section>
  </page>
  <page>
    <page_number>
      204
    </page_number>
    <section>
      <annotation>
        Schmalz. [...] bet:
      </annotation>
      <body>
        Verer [...] dar¬
      </body>
    </section>
  </page>
</document>
```

Xml like document structure modelisation and extraction with DAN

Missions

1- Bibliographical study

The work will start with a review of very recent literature relevant to the subject regarding competitive approaches to DAN, among which the Document Understanding Transformer (DONUT) proposed in [1] will be analyzed more in depth. The bibliography will also cover generative models for sequence generation (XML-like TAGGED text) among which the Generative Adversarial Network should receive much attention.

2- Implementations

The bibliographical study will allow to choose generative models to implement and test on specific tasks. The selected approach should be able to integrate the curriculum learning strategy of the DAN. We should start with the generation of synthetic documents for the known benchmark datasets like RIMES, IAM, READ2018, taking also into account the curriculum strategy.

3- Layout only extraction with DAN

Upon successful development of part 2, the work should continue exploring the proposed approach on new challenging datasets with more complex layout, such as mathematical expression recognition, newspaper reading order extraction etc… In some cases, being able to extract the structure of the document and its reading order should be of particular interest, notably in the case of old newspapers with very large images. Multi-resolution should also be relevant in this respect.

LITIS

LITIS (Laboratoire d'Informatique, Traitement de l'information et des Systèmes) is a research laboratory associated to the University of Rouen Normandie, University of Le Havre Normandie Normandie, and School of Engineering INSA Rouen Normandie. Research at LITIS is organized around 7 research teams which contribute to 3 main application domains. The Machine Learning team of LITIS is developing research in modeling unstructured data (signals, images, text, etc…) with machine learning algorithms and statistical models. For more than two decades it has contributed to the development of reading systems and document image analysis for various applications such as postal automation, business document exchange, digital libraries, etc…

Position to be filled :
Positions: 6-month internship, with possible continuation as a Ph.D student.
Time commitment: Full-time
Duration of the contract: March 1st 2023 – August 31st 2023
Contact: Prof. Thierry Paquet, Thierry.Paquet@univ-rouen.fr
Location: LITIS, Campus du Madrillet, Saint Etienne du Rouvray, www.litislab.eu

Skills :

- General software development and engineering, Python
- Machine Learning, Computer vision, Natural Language Processing
- Ability to work in a team, curious and rigorous spirit

Bibliography :

G. Kim, T. Hong, M. Yim, J. Nam, J. Park, I. Yim, W. Hwang, S. Yun, D. Han, S. Park , OCR-free Document Understanding Transformer, , ECCV, 2022.