

Fine-Tuning LayoutParser for the Analysis of Historical Italian Newspapers

S. Imboden¹ , L. Mattei¹ , G. Marconi¹ , F. Andrucci¹ , A. Gianelli¹ 

¹ CINECA Interuniversity Consortium, Bologna, Italy

Abstract

We present an initiative carried out by the VISIT team at CINECA, in collaboration with the Italian Ministry of Culture (MIC), aimed at fine-tuning the LayoutParser framework to develop an AI model capable of understanding and decomposing newspaper pages. If successful, this effort would enable the large-scale processing of entire years of Italian newspaper issues, offering significant benefits to a wide range of researchers across multiple disciplines.

CCS Concepts

• **Computing methodologies** → **Computer vision; Machine learning applications**; • **Information systems** → **Digital libraries and archives**; • **Applied computing** → **Document management and text processing**

1. Context and Relevance

The digitization of archival and library documents through the scanning of original materials and the creation of digital collections has been underway for some time. The advent of AI has enriched this field, providing OCR tools (e.g. Tesseract, a widely used open-source engine developed by Hewlett-Packard and maintained by Google, capable of recognizing printed text in more than 100 languages); named entity recognition (NER), which automatically identifies and classifies entities in text (such as names of people, organizations, places, and dates) to support advanced search and meta-data generation; and image captioning.

However, processing images from newspapers presents unique challenges. Before OCR can be effective, a newspaper page must be segmented into elements like headlines, subheadings, paragraphs, and images, preserving their hierarchy.

2. LayoutParser

Among the various open-source AI tools designed for this purpose, LayoutParser [SZD*21] has emerged as the most promising. Based on recent technologies, LayoutParser is available at (<https://layout-parser.github.io/>) and provides a modular architecture that can be adapted to various document analysis tasks. One notable use case is the layout recognition of scientific publications, for which it produces particularly strong results. Although a version of LayoutParser specialized for newspaper processing does exist, it has proven suboptimal [LG24] when applied to the specific dataset under our study. Our evaluations indicate that layout recognition quality is highly dependent on the number of columns used in the page design. The best results are obtained with layouts

consisting of six or seven columns, while accuracy degrades significantly when this structure varies, which is a common problem due to the low generalization capability of the underlying model.

Large initiatives like the European Collaborative Cloud for Cultural Heritage (ECCCCH) and national plans such as Italy's Plan for the Digitalisation of Cultural Heritage (PND) promote shared cultural heritage development. In this context, CINECA's VISIT team aims to fine-tune LayoutParser for historical Italian newspapers using images from *Il Resto del Carlino* (1930–1940), a regional paper founded in Bologna in 1885. The data is provided by *Storia e Memoria di Bologna* (<https://www.storiaememoriadibologna.it/>), a public database curated by the Municipality of Bologna.

3. Fine Tuning

Fine-tuning an AI model for a specific application involves two main phases: preparing the training data and training the model. Data preparation is labor-intensive and detailed below. Training continues from the model's existing state and mainly involves defining metrics to evaluate error—i.e., the difference between inference output and expected results. Errors are measured regularly, guiding experts in assessing progress and determining completion. Training data consist of sample images (e.g., newspaper pages) accompanied by the expected "page segmentation" (a set of rectangular areas with associated area types). The area types considered by the model are (1) title, (2) photograph, (3) drawing, or (4) advertisement. Notably, the model does not recognize "paragraph" or "body text" areas, but these can be inferred using standard computational methods.

The outcome is a fine-tuned version of LayoutParser. The most



Figure 1: the first processing stage detects guidelines on images (shown in dark yellow). They are editable (created, deleted, moved with the yellow handles) and are used as snapping references.

demanding part is compiling a large dataset (estimated at 10,000 image-segmentation pairs), where each "segmentation" has been manually created and verified by a human.

One strategy to reduce the required human effort involves an AI-human collaboration approach. Rather than asking the human operator to manually decompose the page into rectangles from scratch, we can use a previous version of LayoutParser to generate an initial output—albeit imperfect—which the human can then correct.

A second strategy we adopted to further accelerate the process is the systematic use of “snap” mechanisms, a feature that attract the mouse cursor to specific anchor points. In our specific case, we observed that text columns or paragraphs within the same column are often (though not always) delimited by vertical or horizontal lines, referred to here as “guidelines”. These lines can be detected in a preliminary step using computer vision techniques and subsequently used as anchor points, thus making the placement of rectangles faster and more accurate.

4. The Application for the Preparation of the Training Data

Recognizing that the preparation of training data would require simultaneous input from multiple human operators, we developed a collaborative web-based application to support the task. The tool can be used from the browser and is designed to track user activity and prevent potential conflicts between contributors.

The application is made of open source components only, the most relevant being Pocketbase (<https://pocketbase.io/>) the database used as backend, SvelteKit (<https://svelte.dev/>) a javascript framework used in the frontend and OpenSeaDragon (<https://openseadragon.github.io/>) a library supporting image streaming and visualization with pan and zoom.

The application provides features to navigate the collections of images, which are grouped by month. Additionally, it tracks the status of each image and the overall work progress. The page for a specific image provides two sets of tools : (1) tools to create/delete/edit the guidelines (2) tools to create/edit/delete the rectangles. In the latter we make heavy use of the snapping feature, both to the guidelines and to the other rectangles, and we also employed *ad hoc* interac-

tion systems to further streamline the work.

New users become proficient in using the tool in about one hour. Annotation time is 5–15 minutes per page. The processing speed depends on the number of paragraphs on the page, which can range from 40 to 130. Another factor influencing the speed is whether the guidelines are correctly recognized automatically or require manual correction.

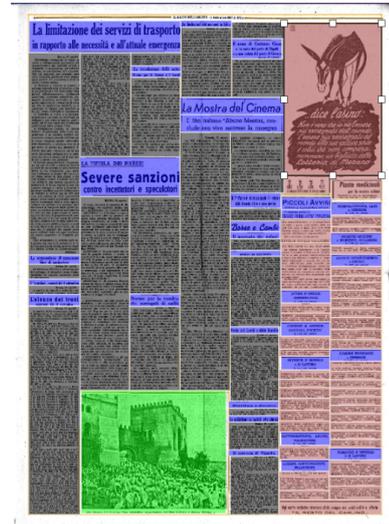


Figure 2: example of an image with a completed annotation. Color represents the type of section. Blue: title, Grey: text, Green: images, Pink: advertisements

5. Conclusions

At the time of writing, we have already completed 1,000 pages and are actively seeking collaborators to help us reach the critical mass necessary to complete this task. Once the dataset is ready, the training will be conducted on Leonardo, CINECA’s high-performance computing infrastructure (<https://leonardo-supercomputer.cineca.eu/>). The expected outcome is a better version of the LayoutParser model specific for newspapers. Several steps can be carried out following layout recognition in order to achieve a meaningful digital transcription of a newspaper’s content. These include linking paragraphs and images to their corresponding headlines, generating a hierarchical description of the page structure, performing Optical Character Recognition (OCR), and applying Named Entity Recognition (NER). However, we are focusing on this particular problem as it represents a bottleneck for the complete process.

We look forward to assessing the outcomes of this initiative.

References

- [LG24] LIGUORI M., GUIDAZZOLI A.: *AI, Cultural Heritage, and Art. Between Research and Creativity. Workshop proceedings – pages 31-40 - February 9-10, 2024*. 12 2024. 1
- [SZD*21] SHEN Z., ZHANG R., DELL M., LEE B. C. G., CARLSON J., LI W.: *Layoutparser: A unified toolkit for deep learning based document image analysis*, 2021. URL: <https://arxiv.org/abs/2103.15348>, arXiv:2103.15348. 1