SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion

Ahmed Nassar[1], **Andres Marafioti**[2], **Matteo Omenetti**[1], **Maksym Lysak**[1], **Nikolaos Livathinos**[1], **Christoph Auer**[1], **Lucas Morin**[1], **Rafael Teixeira de Lima**[1], **Yusik Kim**[1], **A. Said Gurbuz**[1], **Michele Dolfi**[1], **Miquel Farré**[2], **Peter W. J. Staar**[1]

IBM Research[1], HuggingFace[2]

https://huggingface.co/ds4sd/SmolDocling-256M-preview
Version 1.0, March 14th 2025

Abstract

We introduce SmolDocling, an ultra-compact vision-language model targeting end-to-end document conversion. Our model comprehensively processes entire pages by generating DocTags, a new universal markup format that captures all page elements in their full context with location. Unlike existing approaches that rely on large foundational models, or ensemble solutions that rely on handcrafted pipelines of multiple specialized models, SmolDocling offers an end-to-end conversion for accurately capturing content, structure and spatial location of document elements in a 256M parameters vision-language model. SmolDocling exhibits robust performance in correctly reproducing document features such as code listings, tables, equations, charts, lists, and more across a diverse range of document types including business documents, academic papers, technical reports, patents, and forms — significantly extending beyond the commonly observed focus on scientific papers. Additionally, we contribute novel publicly sourced datasets for charts, tables, equations, and code recognition. Experimental results demonstrate that SmolDocling competes with other Vision Language Models that are up to 27 times larger in size, while reducing computational requirements substantially. The model is currently available, datasets will be publicly available soon.

1 Introduction

For decades, converting complex digital documents into a structured, machine-processable format has been a significant technical challenge. This challenge primarily stems from the substantial variability in document layouts and styles, as well as the inherently opaque nature of the widely used PDF format, which is optimized for printing rather than semantic parsing. Intricate layout styles and visually challenging elements such as forms, tables, and complex charts can significantly impact the reading order and general understanding of documents. These problems have driven extensive research and development across multiple domains of computer science. On one hand, sophisticated ensemble systems emerged, which decompose the conversion problem into several sub-tasks (e.g. OCR, layout analysis, table structure recognition, classification) and tackle each sub-task independently. Although such systems can achieve high-quality results for many document types while maintaining relatively low computational demands, they are often difficult to tune and generalize.

On the other hand, in the recent past, great interest has developed around large foundational multimodal models that can solve the whole conversion task in one shot, while simultaneously offering flexible querying and parametrization through prompts. This approach has been made possible by the advent of multimodal pre-training of large vision-language models (LVLMs), which opened up a vast array of opportunities for leveraging diverse data sources, including PDF documents. However, the literature on this topic highlights a significant gap in the availability of high-quality and open-access datasets suitable for training robust multi-modal models for the task of document understanding. Furthermore, relying on LVLMs may introduce common issues associated with such models, including hallucinations and the use of significant computational resources, making them impractical from both a quality and cost perspective.