Multimodal Document Intelligence System

Abstract

This project proposes an end-to-end Document Intelligence system that converts unstructured, multimodal documents—such as scanned pages, PDFs, and handwritten notes—into actionable insights and structured visual representations. At its core is a novel graph-oriented transformer decoder that learns to identify and sequence process steps from document embeddings, then emits node–edge tokens which are rendered as flowcharts. The pipeline begins with OCR and layout analysis (using a vision-language model) to segment and embed text blocks, followed by a retrieval-augmented module for optional question answering and summarization. We will curate a diverse corpus of procedural documents (legal contracts, technical manuals, SOPs, academic protocols) and introduce realistic OCR noise augmentations to improve robustness. Performance will be evaluated on a held-out test set via graph edit distance and node/edge precision–recall, and through a small expert user study to assess clarity and time savings. By unifying process extraction, visualization, and interactive query capabilities in one learned framework—and by open-sourcing all code and data—this work aims to advance the state of the art in multimodal document understanding and automated diagram generation.