OmniParser: A Unified Framework for Text Spotting, Key Information Extraction, and Table Recognition

Jianqiang Wan | Sibo Song | Wenwen Yu |Yuliang Liu | Wenqing Cheng | Fei Huang | Xiang Bai Cong Yao | Zhibo Yang

Alibaba Group| Huazhong University of Science and Technology

Abstract

OmniParser is a unified model designed for visually-situated text parsing (VsTP) tasks, which have recently advanced due to the demand for automated document understanding and the rise of large language models (LLMs) in document question-answering. Unlike previous approaches requiring distinct models for specific tasks, OmniParser can simultaneously perform text spotting, key information extraction (KIE), and table recognition with a single encoder-decoder architecture. Using point-based text generation, OmniParser processes text data in diverse visual formats and achieves state-of-the-art results on multiple datasets for these VsTP tasks, simplifying the previously complex workflows in document parsing.

1. Introduction

The demand for extracting structured information from document images has accelerated research in visually-situated text parsing (VsTP). VsTP methods focus on spotting text within images, recognizing key details, and structuring data in tables, essential for automated document processing. However, past approaches typically used distinct, task-specific models tailored to each VsTP task, leading to more complex systems with limited integration potential. OmniParser aims to unify VsTP tasks under a single model architecture, capable of handling text spotting, KIE, and table recognition simultaneously. Leveraging a two-stage decoding process, it detects and processes text in varied formats, integrating task-specific prompts with structured point-based representations. OmniParser's simplified approach allows it to perform as effectively as separate, specialized models.

2. Related Work

Research in VsTP includes scene text spotting, KIE, and table recognition, each typically requiring a specialized model:

Scene Text Spotting: The objective is to locate and recognize text in various orientations and styles. Transformer-based models have provided improved flexibility, although limitations remain in handling complex layouts.

Key Information Extraction (KIE): Divided into OCR-dependent and OCR-free models, KIE models use different strategies to structure text and entity relationships within document images. OCR-free models, in particular, aim to streamline workflows.

Table Recognition: Table recognition, which includes detecting table structure and extracting cell content, generally uses separate models for structure and content recognition. Few models unify these tasks into a single model. OmniParser integrates these VsTP tasks, utilizing strengths from prior approaches while maintaining simplicity and unified functionality.

3. Methodology

3.1 Unified Model Structure

OmniParser uses a shared encoder-decoder framework that efficiently handles VsTP tasks. This architecture processes text spotting, KIE, and table recognition with a unified input-output system, incorporating task-specific prompts for each task. The model adopts a point-based structured points sequence, simplifying the complexity associated with multiple task-specific configurations.

3.2 Structured Points Sequence

OmniParser's structured points sequence approach encodes positional and task-specific tags within each document image. For example, tokens like `<address>` or `<tr>` are used in KIE and table recognition, respectively, helping the model differentiate tasks without modifying its core structure.

3.3 Two-Stage Decoding

OmniParser's two-stage decoding approach is key to its efficiency and performance:

Stage 1: The model generates center points of text or cells using image embeddings and task prompts.

Stage 2: Based on the center points, OmniParser extracts content and structure, forming complete textual or tabular information for output.

By separating location and content generation, this model achieves better performance in handling complex, diverse document structures.

3.4 Pre-training Techniques

OmniParser applies two pre-training techniques, spatial-aware and content-aware prompting, to enhance its ability to process structured document elements:
Spatial-aware prompting: Helps the model focus on specific spatial regions in documents.
Content-aware prompting: Guides the model in recognizing content, aiding in tasks like entity extraction.
These methods optimize the model's spatial and semantic understanding, improving performance across all VsTP tasks.

4. Experiments

4.1 Implementation Details

OmniParser was pre-trained on text-rich datasets with a range of spatial and content-based augmentations. Fine-tuning was done separately for each VsTP task using task-specific datasets and evaluation metrics.

4.2 Text Spotting

OmniParser's performance was evaluated using three popular scene text datasets: Total-Text, ICDAR 2015, and CTW1500. These datasets are designed to test arbitrary-shaped text detection and spotting in various document layouts. On these datasets, OmniParser achieved high end-to-end recognition results under different lexicon settings, demonstrating its ability to accurately recognize complex text layouts without additional models.

4.3 Key Information Extraction

For the KIE task, OmniParser was tested on the CORD and SROIE datasets. These datasets include diverse document structures with predefined entity fields, such as company names, dates, and addresses. Evaluation metrics included field-level F1 scores and tree-edit-distance-based accuracy. OmniParser outperformed prior generation-based methods, achieving high accuracy across both datasets, highlighting its flexibility in identifying key information across varying layouts.

4.4 Table Recognition

OmniParser's table recognition capabilities were assessed on PubTabNet and FinTabNet. The model demonstrated strong performance in recognizing both the structure and content of tables using Tree-Edit-Distance-based Similarity (TEDS). OmniParser's point-based approach simplified processing and improved accuracy, particularly in identifying complex table layouts.

5. Analysis

5.1 Pre-training Strategy Analysis

Through ablation studies, OmniParser's pre-training strategies of spatial and prefix-window prompting were shown to improve accuracy and generalization. Spatial prompting enhanced the model's perception of text coordinates, while prefix prompting improved its understanding of text relationships.

5.2 Architectural Variants

Different architectural configurations, including the encoder-decoder setups, were tested to maximize performance. Swin-B outperformed ResNet50 in handling VsTP tasks, and unshared decoders improved accuracy across text spotting tasks, demonstrating that certain architectural choices can optimize multi-task parsing.

5.3 Decoder Length Analysis

Experimentation with decoder lengths highlighted the balance between sequence length and efficiency. Optimized lengths provided enhanced accuracy in table recognition while maintaining processing speed.

6. Conclusion and Future Work

OmniParser is an effective, unified approach for VsTP tasks, streamlining text spotting, KIE, and table recognition within a single model. The two-stage decoding process, supported by structured points, has allowed OmniParser to achieve high accuracy across tasks. Future work will focus on expanding the model's capabilities to incorporate more complex document elements, such as images and charts, and to enhance real-world adaptability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62225603) and the Alibaba Innovative Research (AIR) program.