Multimodal Information Extraction of Supermarket Leaflets

Xincheng Liao, Junwen Duan*, Yixi Huang, Jianxin Wang
Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering,
Central South University, Changsha, Hunan, China
{ostars, jwduan, yx.huang}@csu.edu.cn, jxwang@mail.csu.edu.cn
https://github.com/OStars/RUIE

Abstract

TODO

1 Introduction

1.1 Motivation

1.2 Problem Statement

1.3 Contributions

2 Related Works

Deal Detection. **Optical Character Recognition**. OCR-Model-Driven methods use OCR tools to acquire text and bounding box information. Subsequently, they rely on the models to integrate text, layout, and visual data.

Information Extraction.

3 Deal Detection

3.1 Datasets

3.2 Model

3.3 Experiments

4 Information Extraction of Supermarket Deals

Subsequently to the detection and separation of supermarket deals, the overarching goal of using these in applications requires the extraction of the various useful information from these deals. The task of extracting meaningful and structured information from supermarket deal images is a complex and multi-faceted problem, requiring the integration of various modalities, overcoming challenges posed by noisy data, and leveraging deep learning techniques. This section provides a detailed exploration of the underlying challenges, methodologies, and the theoretical and practical framework needed to perform such extraction.

* Corresponding author. Email: jwduan@csu.edu.cn

4.1 Challenges in Information Extraction for Supermarket Deals

Supermarket deal extraction involves a series of challenges that require careful attention and sophisticated methods to resolve. These challenges include the diversity of layout and format in the images, the multimodal nature of the data, the frequent occurrence of Optical Character Recognition (OCR) errors, and the specific linguistic and cultural issues posed by the German language.

High Variety in Layouts and Visual Elements. Supermarket deals exhibit considerable variability in layout, color schemes, font choices, and text sizes, which hinders automatic detection and extraction. Deals are often printed with superscripted decimals or without clear separation between integer and fractional components of the price (e.g., "2.29" might appear as "229"). This phenomenon exacerbates OCR difficulties. Additionally, original prices may be struck through, while deal prices are sometimes printed in drastically different fonts than those the OCR model was trained on. Furthermore, overlapping elements—such as product images or additional textual information—often interfere with text extraction and localization. These diverse visual elements require robust and flexible models capable of accommodating such variability.

Multimodal Information. Supermarket deal images consist of both visual (e.g., product images) and textual (e.g., brand names, prices, and product descriptions) modalities, each contributing different types of information. While text provides rich information about the product's identity, pricing, and units, the image serves as a supplementary modality that aids in product identification, brand recognition, and other visual cues. Multimodal fusion, where information from both text and image domains is integrated, is a key challenge, and existing models must effectively leverage both sources to provide high-quality extraction.

Conversion Errors and Linguistic Challenges. OCR systems frequently introduce errors, particularly when dealing with non-standard fonts, noisy backgrounds, or small text. This is particularly problematic when extracting numerical data such as prices. Common errors include misinterpretation of decimals and digits (e.g., "229" instead of "2.29") and missing characters. To address this, OCR post-processing steps, such as error correction using context-based reasoning or dedicated error models, are required. Additionally, the presence of proper nouns—particularly brand names—adds complexity, as these entities may not appear in typical language models, making their extraction difficult. Furthermore, the German language presents its own set of challenges, including compound words, specific punctuation conventions, and diverse word forms, all of which must be accounted for in any robust extraction model.

Ambiguities. The presence of overlapping elements—whether textual or graphical—can lead to a degradation in the extraction accuracy. This challenge is compounded when the overlap involves essential information, such as when the original price is partially obscured by a product image or strike-through marks. Furthermore, ambiguity in labeling and the context in which different pieces of information appear in the deal image can create difficulties in associating text with the correct object (e.g., associating a price with a product rather than the surrounding descriptive text).

4.2 Formal Approach

Formally, the task of information extraction from a deal image can be defined as the identification of specific entities, including the product name, brand, original price, deal price, and unit. We define the following set of output labels:

$$\mathcal{Y} = \{y_i\}_{i=1}^{n}, \tag{1}$$

where y_i represents the i-th entity in the deal image. In the following, the entities are defined as:

- $y_{\text{product\_name}}$: The name of the product.
- y_{brand}: The brand of the product.
- $y_{\text{original\_price}}$: The original price of the product.
- $y_{\text{deal\_price}}$: The deal price of the product.
- y_{unit}: The unit of the product (e.g., weight, volume).

To extract these entities, one aims to learn a function $f(\cdot)$ that maps an input image I to the desired outputs:

$$f : I \rightarrow \mathcal{Y}, \tag{2}$$

where I is the deal image, and $\mathcal{Y}$ represents the extracted entities.

The choice of the function $f(\cdot)$ is an architectural decision that depends on the specific requirements of the task, the nature of the data, the available resources, and the desired performance metrics.

- Approaches - OCR + LLM - Vision Encoder Decoder Model - Donut - LVLM

4.3 Optical Character Recognition

4.4 Experiments

Setup. - Common OCR solutions: Tesseract, EasyOCR, PaddleOCR, DocTR - Goal: Which OCR solution is the best for our task? -

4.5 Evaluation

- LVLM, OCR+LLM, Donut - OCR + LLM Vergleich

5 Application

6 Database

7 Front End / Webapp

8 Conclusion

8.1 Summary

8.2 Future Work

ww