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

Automated information extraction from supermarket leaflets poses significant challenges due to the unstructured nature of promotional content, diverse layouts, and varying text-image relationships. The ability to accurately extract product names, prices, and discount details is critical for applications such as competitive analysis, dynamic pricing strategies, and digital retail transformation. Traditional OCR-based approaches often struggle with these complexities, as they are not well-suited for handling visually complex layouts that integrate both textual and graphical elements. Consequently, there is a growing need for advanced methodologies that leverage deep learning and multimodal techniques to improve extraction accuracy and robustness.

1.1 Motivation

Retailers distribute promotional leaflets regularly, providing crucial information about product pricing, discounts, and special offers. Extracting structured data from these leaflets is essential for a variety of downstream applications, including price monitoring, personalized marketing, and retail analytics. However, the nature of these documents presents several challenges. Leaflets often feature intricate layouts that combine product images with multiple text regions, making it difficult to separate relevant information from decorative or promotional elements. Furthermore, text appears in different font styles, orientations, and colors, increasing the difficulty of standard OCR-based extraction. The presence of multilingual content and retail-specific terminologies further complicates the process, as many OCR systems struggle to recognize domain-specific jargon or currency symbols correctly. Additionally, supermarket leaflets often contain noisy backgrounds, overlapping elements, and handwritten annotations, all of which contribute to OCR errors. Given these challenges, traditional rule-based and template-based methods fall short of delivering reliable results. Recent advancements in deep learning, particularly in vision-language models, provide a more adaptable and powerful approach to structured information extraction. Large Vision-Language Models (LVLMs) and end-to-end document understanding architectures such as Donut have demonstrated significant potential in tackling these challenges by integrating contextual understanding with visual processing.

1.2 Problem Statement

The core problem addressed in this work is the extraction of structured product and pricing information from supermarket leaflets. Given an image I containing both textual and visual information, the objective is to extract a structured representation $\mathcal{D}$ that encapsulates relevant entities. More formally, this representation can be defined as:

$$\mathcal{D} = \{(p_i, v_i, c_i)\}_{i=1}^{N}, \tag{1}$$

where p_i represents a product name, v_i denotes its corresponding price, and c_i categorizes the type of promotional offer, such as a direct discount or a multi-buy deal. Several challenges make this task particularly difficult. First, segmenting and localizing relevant text and graphical regions in a cluttered layout is non-trivial, as different sections of the leaflet may contain multiple, overlapping elements. Additionally, associating product names with the correct prices requires context-aware processing, particularly when multiple prices exist for a single product due to variations such as original price, discounted price, and special membership discounts. Finally, OCR errors, layout inconsistencies, and missing textual fields introduce additional obstacles, necessitating robust extraction techniques that

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

can generalize across different leaflet designs and promotional structures.

1.3 Contributions

This work makes the following key contributions:

- We conduct a systematic evaluation of various information extraction approaches, including traditional OCR pipelines, OCR-LLM hybrid models, and end-to-end deep learning frameworks.
- We introduce a normalization pipeline to improve OCR output consistency, mitigating common recognition errors in promotional leaflets.
- We benchmark state-of-the-art document understanding models such as Donut and LVLMs on supermarket leaflet datasets, analyzing their effectiveness in structured information extraction.
- We provide an open-source implementation and dataset annotations to facilitate future research in retail document analysis.

These contributions aim to bridge the gap between traditional OCR-based methods and modern deep learning approaches, advancing the field of automated leaflet analysis.

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 Nature of Supermarket Leaflet Data

Supermarket leaflets constitute a complex and heterogeneous data source, characterized by multimodal content with varying levels of structured and unstructured information. This data presents unique challenges in computational processing due to its inherent ambiguities, multi-instance object representation, and spatial dependencies. The reader is encouraged to simulanously inspect the visual representation of the specific data-related challenges in Figure 1 while reading ahead.

Formally, let $I \in \mathbb{R}^{H \times W \times C}$ denote a supermarket leaflet image, where H and W represent height and width, and C indicates the number of channels. The primary goal in leaflet analysis is to extract a structured representation $\mathcal{D} = \{(c_i, p_i, d_i, m_i)\}_{i=1}^{N}$, where each tuple represents category c_i, price p_i, product description d_i, and brand marker m_i for N detected items.

The ambiguity in leaflet data stems from multiple contributing factors. First, price recognition is hindered by the existence of multiple price points associated with a single product. Given product P_j, we define a set of associated prices $P_j = \{p_{j,1}, p_{j,2}, ..., p_{j,K}\}$, where K denotes the number of distinct price entries, including original prices, discounts, membership-based reductions, and additional incentives (e.g., coupon-based deductions). The decision function for selecting the correct price entry can be formulated as:

$$p_j^* = \arg\max_{p \in P_j} \Phi(p, C), \tag{2}$$

where $\Phi(p, C)$ represents a learned scoring function incorporating context C, including proximity to textual cues and visual emphasis features.

Furthermore, the representation of products in supermarket leaflets does not conform to a strict one-to-one mapping between textual and visual elements. Some product deals encapsulate multiple distinct items, leading to a one-to-many mapping between product images and extracted textual features. Let $X = \{x_i\}_{i=1}^{M}$ represent a set of detected product images and $T = \{t_j\}_{j=1}^{N}$ a set of textual entities; the optimal pairing function $\mathcal{M} : X \to 2^T$ requires the resolution of occlusions, segmentation artifacts, and inconsistent typographic hierarchies.

In the context of visual information extraction, a major challenge arises from the misinterpretation of numerical values due to errors in punctuation extraction. Specifically, the process is prone to systematic errors in decimal place recognition due to font-specific artifacts and noisy backgrounds. Let s_t denote the OCR-extracted string, where $s_t \sim P(S)$ represents the probability distribution over possible transcriptions. A corrective decoding function $\Gamma(s_t)$ can be optimized via:

$$\Gamma(s_t) = \arg\max_{s \in \mathcal{S}} P(s | s_t, \mathcal{C}), \tag{3}$$

where $\mathcal{S}$ is the space of plausible price values, and $\mathcal{C}$ integrates surrounding contextual information.

Finally, spatial dependencies play a critical role in entity association. Traditional object detection

methods relying purely on bounding box regression, such as Faster R-CNN (**?**), may be insufficient due to the lack of explicit spatial reasoning mechanisms. Therefore, the incorporation of spatial modeling is essential to ensure accurate entity association and disambiguation.

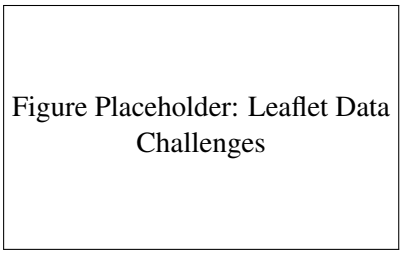

Figure 1: Visual representation of the challenges in supermarket leaflet data.

4 Deal Detection

4.1 Datasets

4.2 Model

4.3 Experiments

5 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.

5.1 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).

5.2 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:

$$\mathcal{Y} = \{y_i\}_{i=1}^{n}, \tag{4}$$

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

- $y_{\texttt{product\_name}}$: The name of the product.
- $y_{\texttt{brand}}$: The brand of the product.
- $y_{\texttt{original\_price}}$: The original price of the product.
- $y_{\texttt{deal\_price}}$: The deal price of the product.
- $y_{\texttt{unit}}$: 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:

$$f : I \rightarrow \mathcal{Y}, \tag{5}$$

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.

5.3 Architectural Approaches to Information Extraction

Among the various existing architectural approaches to information extraction, several methodologies have been developed to address the inherent complexities in supermarket deal images. These methods can be categorized into classical approaches, multi-stage traditional models, hybrid OCR-LLM systems, and end-to-end deep learning frameworks. Each paradigm offers distinct advantages, and their applicability depends on computational resources, dataset characteristics, and performance requirements.

5.4 Traditional Multi-Stage Methods

Traditional architectures decompose the problem into sequential sub-tasks and solve these independently through sub-task-specific systems. While these methods have been studied extensively and with high variability in methodology, they are often divided into three main stages: text detection, text recognition, and key-value pair extraction.

Text Detection. Text detection is the process of identifying regions in the image that contain textual information. Commonly used methods include rule-based systems, as well as mostly small object detection neural networks to detect regions of textural interest by predicting M bounding boxes $\{b_i\}_{i=1}^{M}$.

Text Recognition. Each bounding box b_i is used to extract the corresponding region in the image, which is passed to a text recognition model to convert the visual representation into a textual one. Since this task is often more complex than it might appear due to decoding errors, there exists a high variety of models that can be used for this task, such as CRNN (**?**) or transformer-based recognizers. Given a set of M bounding boxes $\{b_i\}_{i=1}^{M}$, the goal is to predict the corresponding textual entities $\{t_i\}_{i=1}^{M}$.

$$\{t_i\}_{i=1}^{M} = \arg\max_{\{t_i\}_{i=1}^{M}} P(\{t_i\}_{i=1}^{M} | \{b_i\}_{i=1}^{M}). \tag{6}$$

Text detection and recognition, together with pre-and post-processing steps, is often referred to and standardized as Optical Character Recognition (OCR).

Key-Value Pair Extraction: Finally, extracted text is structured into meaningful entities. Named Entity Recognition (NER) models or Graph Neural Networks (GNNs) can be used to learn associations between extracted elements:

$$\mathcal{A}^* = \arg\max_{\mathcal{A} \subseteq E} \sum_{(i,j) \in \mathcal{A}} \Psi(v_i, v_j), \tag{7}$$

where E represents textual adjacency relationships.

While these types of appraoches are widely used, they inherit a high complexity due to the need for multiple models with each being trained, evaluated, tested and managed independently, which

also leads to a higher computational overhead due to input-output transformations between the stages. Furthermore, the lack of end-to-end training can lead to suboptimal performance and error propagation, which is often seen when applying these types of models to edge cases or unseen data.

5.5 Hybrid OCR-LLM Models

The rise of large language models (LLMs) has led to the development of hybrid OCR-LLM models that combine the strengths of OCR tools with the linguistic capabilities of LLMs. While these appraoches still rely on a more traditional process to convert visually appearing text into a machine-readable form, the LLM is used to enhance the robustness of the OCR output by providing contextual reasoning capabilities.

$$\hat{y} = F_\theta(T), \tag{8}$$

where T is the OCR output and F_θ is a pretrained, general-purpose LLM that is being prompted with a specific task.

5.6 End-to-End Models

The

End-to-End (E2E) approaches eliminate pipeline fragmentation by learning a unified mapping $F : I \to \mathcal{D}$. Two major architectures dominate this space:

Vision Encoder-Decoder Models: Models such as Donut (**?**) encode input images via a Vision Transformer (ViT) and generate structured outputs via an autoregressive decoder:

$$P(Y|I) = \prod_{t=1}^{T} P(y_t|I, y_{<t}). \tag{9}$$

These methods improve robustness by directly training on image-text pairs without explicit OCR supervision.

Large Vision-Language Models (LVLMs): Multimodal transformers such as BLIP (**?**) and LLaVA (**?**) leverage joint vision-text embeddings to infer structured outputs. Given an image-text embedding $z = f_\theta(I, T)$, task-specific decoding is performed via:

$$\hat{y} = \text{Decoder}(z), \tag{10}$$

where Decoder is a domain-adapted transformer head.

E2E models significantly reduce error propagation but require large-scale annotated datasets for effective training. Their adoption is growing with advancements in multimodal pretraining strategies and self-supervised learning.

5.7 Dataset Creation

Since the availability of German supermarket leaflet data is literally not findable, a custom dataset was created by manually annotating a collection of supermarket deal images. The dataset includes images from the most common supermarket chains, each with unique layout, fonts, and colors to, in general, ensure the highest diversity and robustness possible. Each image is annotated with a corresponding label file that includes the desired entities as well as the image$_id$.$The dataset is split into a training and a validation set fo$

Table 1: Leaflet-IE Dataset

Entity	Sample Size
`image_id`	372
`brand`	357
`product_name`	370
`original_price`	286
`deal_price`	372
`weight`	369

Even though the Leaflet-IE dataset is relatively small, the reader will be able to see that the dataset is sufficient to evaluate the performance of different IE appraoches as well as to train an competitive end-to-end model. However, as the reader may notice, the dataset is solely focused on single product deals at this point and explicitly does not include

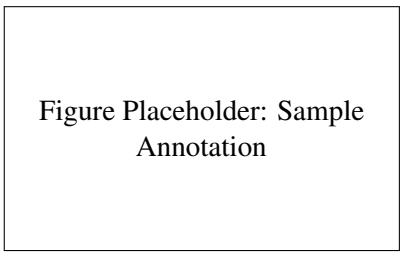

Figure 2: Sample images with annotations from the custom dataset.

6 Experiments

The primary objective of the experiments is to identify the most effective approach for extracting structured information from supermarket leaflets. The

evaluation considers general performance trends and use cases to determine candidate methodologies, including OCR + LLM, Donut, and LVLM.

6.1 Implementation Details

Hardware Configuration: The experiments were conducted on two distinct computational setups:

- Station 1: NVIDIA RTX 4080 GPU
- Station 2: NVIDIA GTX 1080 GPU

Software Stack: The following libraries and frameworks were utilized:

- PyTorch
- HuggingFace Transformers
- OpenCV

6.2 Comparison of Hybrid OCR+LLM, Donut, and LVLM

- Comparison of Hybrid OCR+LLM, Donut, and LVLM - Evaluation on Leaflet-IE validation subset. - Evaluation Metrics: Accuracy, levenshtein distance

- Sections X and Y show prepended studies to determine the best model in a selected candidate pool for OCR and LLM.

Results. TODO

6.3 Plain Deal OCR Performance

The baseline experiment focuses on the effectiveness of OCR models in extracting values directly from the document images.

Models Evaluated:

- Tesseract
- EasyOCR
- PaddleOCR
- DocTR

- Data: All 372 samples of the Leaflet-IE dataset were used for evaluation since all OCR models probably have not seen the data before.

Evaluation Metrics:

- **Accuracy**: Measures whether the expected value is present in the OCR output.
- **N-gram Accuracy**: Determines if an n-gram substring of the expected value exists in the OCR output, computed for $n \in \left[\frac{|v|}{2}, |v| - 1\right]$, where $|v|$ represents the string length.

To improve comparability, OCR outputs were evaluated in different levels of normalization. Preprocessing function

Results. TODO

The performance comparison across different OCR models is visualized in Figure 3, where each system's accuracy is evaluated across different deal structures. The raw and normalized outputs provide insights into the robustness of each approach in handling textual inconsistencies.

Figure Placeholder: OCR Model Performance

Figure 3: OCR Model Performance on Plain Deal Text Extraction

6.3.1 OCR + LLM Performance

Experiment Question: -Which LLM is best suited to work with OCR outputs? -Which LLM + OCR combination works best in IE?

Data: - All 372 samples of the Leaflet-IE dataset were used for evaluation since all OCR models as well as the LLMs probably have not seen the data before.

The LLM is tasked via a prompt to extract the entities from the OCR output.

Models Evaluated: - "llama3.1$_8b$" : "$Llama3.1[8b, Q4]$", −"$qwen2.5_1.5b$ − $instruct$ − $q8_0$" : "$Qwen2.5[1.5b, Q8]$", −"$llama3.2_3b$ − $instruct$ − $q8_0$" : "$Llama3.2[3b, Q8]$", −"$qwen2.5_7b$" : "$Qwen2.5[7b, Q4]$"

Evaluation Metrics: - For each OCR: Per Entity accuracy and levenshtein distance between predictio

Normalization: The reader might get confused with the normalization of the OCR experiment. Here, the decision was made to keep the OCR results as they are, as normalization may remove valuable information. In addition, the LLMs are

considered to deal with such inconsistencies and errors as input.

Here, to improve comparability, The entity values for the comparison between the LLM prediction and the ground truth value were compared in raw form as well as with a normalization prepended. The normalization function used is as follows: . The normalization procedure applied character replacements and filtering based on linguistic artifacts commonly encountered in OCR errors.

Results.

6.4 Donut Fine-Tuning Performance

Experiment Question: - Is it possible to train a Vision Encoder Decoder model to extract end2end information from supermarket leaflet deal images?

7 Application

7.1 Database

7.2 Evaluation

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

8 Application

9 Database

10 Front End / Webapp

11 Conclusion

11.1 Summary

11.2 Future Work

ww