November 2023

Automatic Structured Menu Extraction from Menu Photographs

Bo Lin

Jinyang Yu

Jorge Nario

Xinru Yang

Automatic Structured Menu Extraction from Menu Photographs

ABSTRACT

Restaurant menu images can be utilized to automatically obtain structured data about dish names, prices, etc. However, the raw optical character recognition (OCR) output suffers from low quality and OCR techniques do not have sufficient ability to adapt to the diversity in language and design of restaurant menus. A language model can be used together with OCR to identify dish names and other content through a named entity recognition (NER) process. However, this is not scalable due to the requirement of a large, labeled dataset across languages and countries. This disclosure describes the use of a multimodal large language model (LLM) to automatically generate digital structured menus from restaurant menu photographs. The use of a multimodal large language model enables automatic creation of structured digital menus that include price, description, ingredients, etc. without the requirement of a large amount of labeled data and can also overcome difficulties associated with low quality photographs. The capabilities of multimodal LLMs are leveraged by formulating the task of menu understanding from the user-provided photos as a multimodal information extraction or a visual question answering task which fits naturally with the framework of multimodal pretrained large models.

KEYWORDS

- Large language model (LLM)
- Multimodal LLM
- Named entity recognition
- Menu extraction
- Menu understanding
- Menu photograph

BACKGROUND

Digital maps, restaurant reviews, online ordering/delivery services, etc. host images related to topics such as food, restaurants, menus, etc. One of the popular types of photographs is photographs of restaurant menus.

Many restaurant menus are available only in physical format. Even if a menu is available in digital format, it may not be easily navigable. For example, information such as prices, ingredients of individual food items, nutrition information, etc. may not be available in a digital menu. Menu photographs can be used to extract and provide information about dishes available at different places in a digital format to assist users. Such information can include dish names, prices, ingredients, etc. and can help users in selecting restaurants and in ordering.

Automatic extraction of information menu photographs is challenging. Menu photographs provided in various ways (e.g., user-uploaded, restaurant-provided) are often of mixed quality, e.g., blurred, captured under poor lighting, captured at awkward shooting angles, etc. Also, menus are diverse across languages and cultures. These issues make automatically extracting information from menus challenging. For example, one of the approaches to extract menu information is to perform optical character recognition (OCR) on a menu photograph to extract text and identify the common menu items against a dictionary or knowledge base. OCR quality degrades with the input photograph quality, and it may not work well across different languages and fonts. Another approach is to use OCR along with a language model to process the text to identify menu names through a named entity recognition (NER) process. However, these techniques may not be scalable due the requirement of large amounts of labeled data and may not be able to provide a structured menu that includes price, description, ingredients, etc.

DESCRIPTION

This disclosure describes the use of a multimodal large language model (LLM) to automatically generate digital structured menus from restaurant menu photographs. The use of a multimodal large language model enables automatic creation of structured digital menus that include price, description, ingredients, etc. without the requirement of labeled data and can also overcome difficulties associated with low quality photographs.

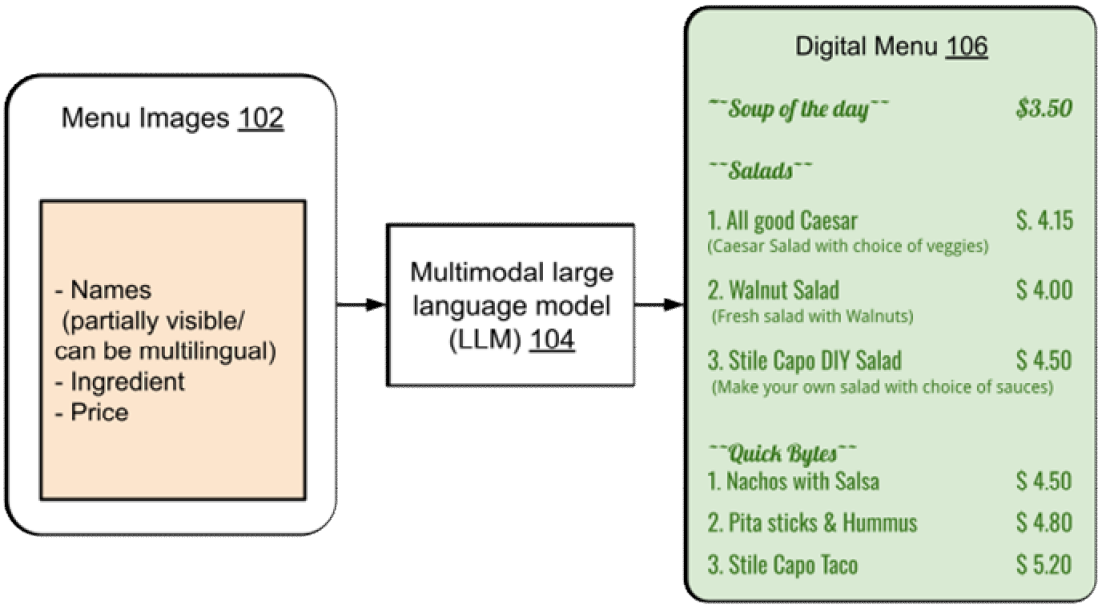

Fig. 1: Generating a structured menu from menu photographs

Fig. 1 illustrates an example of structured menu generation from images of menus. Restaurant menu images (102) are obtained. The images may contain multilingual information (e.g., dish names in Hindi and prices in Roman numerals), or the text may not be readable in certain cases. Menu images are fed to a multimodal large language model (LLM) (104). The LLM is tasked with extraction text from images to generate a structured digital menu (106).

The use of multimodal pre-trained large models can help in the complex task of menu extraction. The capabilities of multimodal LLMs are leveraged by formulating the task of menu

understanding from the photos (including user-provided photos, as well as photos obtained via other permitted sources) as a multimodal information extraction or a visual question answering task which fits naturally with the framework of multimodal pretrained large models. Further, model capabilities in handling multilingual, multicultural data are leveraged in extraction of the content. This approach is substantially simpler to set up. For example, an instruction-tuned multimodal large model and a set of prompts can be utilized. Further, LLM-based menu extraction can scale well across subtasks (e.g., extracting prices, ingredients, nutritional information, etc.) and the described techniques can scale well across languages. The use of an LLM also eliminates the need for very costly massive data labeling.

In situations where the text extraction is partial, appropriate text such as names with correct description, may be generated by the LLM (and verified) to complete the menu. Translated versions of the menu can automatically be provided once the structured digital menu is created. Menu information can be provided in different applications such as digital maps, online ordering/delivery services, search engines, etc.

CONCLUSION

This disclosure describes the use of a multimodal large language model (LLM) to automatically generate digital structured menus from restaurant menu photographs. The use of a multimodal large language model enables automatic creation of structured digital menus that include price, description, ingredients, etc. without the requirement of a large amount of labeled data and can also overcome difficulties associated with low quality photographs. The capabilities of multimodal LLMs are leveraged by formulating the task of menu understanding from the user-provided photos as a multimodal information extraction or a visual question answering task which fits naturally with the framework of multimodal pretrained large models.