An Image-Based Classifier for Text Extraction from Restaurant Menus

Sandeep Kalari
skala003@odu.edu

Sai Chandhar Annapragada
sanna002@odu.edu

***Abstract*— A menu is a collection of items that an organization or industry displays to the public. It is of various types where it can be used for a record as an index and to display various options/products that are offered in organizations such as hospitals, restaurants, and so on.In the restaurant industry, the menu is primarily used as a communication tool. In the current scenario, it is most likely used as images and pdf with a list of food items classified into multiple sections such as appetizers, main courses, beverages, and so on, and with their corresponding details. The list on a few Menu cards would be quite long, making it difficult for partially sighted people to check the object that they are interested in. So, in this scenario, digitizing menu cards into a simpler tree type structured web page can help all consumers and partially sighted people to quickly navigate to the items that they are interested in. Furthermore, the interesting feature that can be used in implementing this project is for visually impaired persons to access the menu cards that are in image format, which would not be possible if the Menu is only in image format because we cannot convert to speech form directly from an image, but here in this project, in addition to pdfs, we would be converting the images data to the webpage layout as well, so this would help them by converting the content on the webpage to speech, they can access the Menu card in image format. .**

***Index Terms*— OCR,,Donut-OCR,MMOcr(pytorch based),Unsupervised Learning,Data Extraction**

I. Current Research and drawbacks

The main objective of this project involves extracting data from images.Current research gives the data from images using OCR.Working with documents for OCR can be difficult in a variety of ways. The key challenges are the source of input data and its attributes. Images of documents captured on mobile phones or handheld devices may be distorted. This is frequently caused by the method of capture, lighting issues, cluttered backgrounds such as watermarks or design patterns, low print quality, flash glare when captured in low-light conditions, if the document was printed on glossy material, variation in resolutions, and so on. The OCR engine becomes confused and produces errors if the input document is skewed.

The Current approach for extracting data from images is done with the following way a.reading the texts that are displayed in the image of the document b.comprehensive comprehension of the text of the document.

The current OCR architecture is shown in the following figure

Another intriguing aspect of our project that we concluded was, "Once the data is collected from the image, we need to create a model that will automatically recognize and categorize the data." For instance, when we scan a hotel menu, it need to instantly identify the item's category. So that a blind person or someone who needs physical assistance can understand the context of the menu by just using a voice assistant.

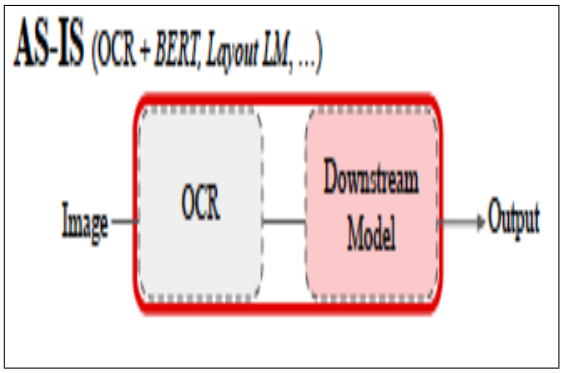

Fig. 1. Barnes-Hut domain decomposition

As a result, the current research uses a few libraries to extract the data, but linking to a voice assistant will provide a new dimension to this research.

II. Our Solution to the Problem

Since our project involves mainly on image extraction we would like to build a model using unsupervised learning to perform semantic way to extract features from images.

A particularly adaptable method for locating areas of interest inside images is provided by object detection models like MaskRCNN and its forerunners. As you may expect, object detection and non-traditional OCR are closely related. The only classes of objects that matter in this situation are text objects and everything else. With this viewpoint, we can train a model that is very similar to MaskRCNN to find text-rich regions of interest (RoI) in an image, often known as text localization.

We go above and beyond the standard framework by modeling a direct mapping from a raw input image to the desired output. This is done without the use of optical character recognition (OCR). We present a new VDU model that does not use OCR in order to address the problems that have arisen as a result of the dependence on OCR. Our model is based on a Transformer-only architecture, and this architecture is referred to as the Document understanding transformer. This architecture was developed as a direct result of the enormous amount of success that was observed in vision

and language (Donut). As part of the minimal baseline that we have developed, we will present a straightforward architecture as well as a technique for pre-training. Figure demonstrates that in spite of its apparent lack of complexity, Donut is capable of achieving an overall performance that is either on par with or superior to that of earlier approaches. A pre-train-and-fine-tune scheme is what we use for the Donut training [8,63]. During the pre-training phase, Donut is instructed on how to read the texts by predicting the next words by conditioning jointly on the image and previous text contexts. This is done in order for Donut to become proficient in reading the texts.

Donut will be able to read the texts more effectively as a result of this. The document images and the text annotations that were associated with them were used to train Donut before it was deployed. Because we have a clear objective for the pretraining phase (namely, reading the texts), we can easily achieve domain and language flexibility by utilizing synthetic data during the pretraining phase. This is the case because our goal for this phase is so straightforward. During the stage of fine-tuning, Donut acquires the knowledge necessary to comprehend the entirety of the document in accordance with the task that lies further downstream. We were able to demonstrate that Donut possesses a powerful understanding ability by conducting in-depth evaluations on a wide variety of VDU tasks and datasets.

This allowed us to demonstrate that Donut possesses a powerful understanding ability. The findings of the experiments indicate that a basic VDU model that does not use OCR is able to achieve state-of-the-art performance in terms of both speed and accuracy. The following is a summary of the contributions in a more concise form: 1. Instead of using optical character recognition (OCR), we propose a brand new method for working with VDUs. To the best of our knowledge, this technique is the first one that we are aware of that employs an OCR-free Transformer that is trained in an end-to-end manner, and we are calling it the "first method." 2. Here, we present a straightforward pre-training method that enables the use of fabricated data while still maintaining its accuracy. By utilizing our generator SynthDoG, we demonstrate that Donut can be easily extended to a setting that supports multiple languages.

In contrast to this, the conventional approaches necessitate the retraining of a commercially available OCR engine, and as a result, they are not applicable within the parameters of this discussion. 3. We perform extensive experiments and analyses on both public benchmarks and private industrial datasets, and the results demonstrate that the proposed method not only achieves state-of-the-art performances on benchmarks, but it also has many practical advantages (such as being cost-effective) in real-world applications. 4. The results demonstrate that the proposed method not only achieves state-of-the-art performances on benchmarks, but it also has many practical advantages (such as being scalable) in real-world applications.

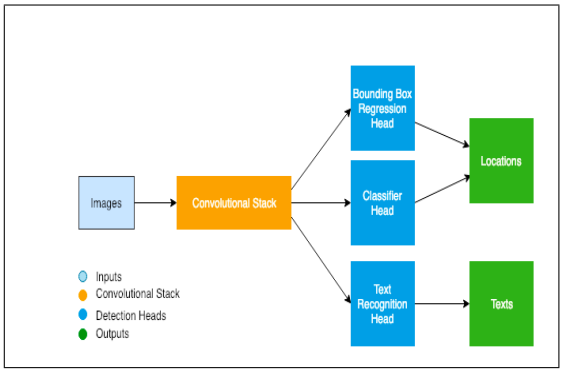

Fig. 2. Barnes-Hut domain decomposition

III. Related Work

Recent tendencies in OCR research have focused on employing deep learning models in both of the field's two substeps: 1) a detector is used to anticipate the location of text areas; 2) a text recognizer is then used to identify each character present in the cropped image instances. Both are trained with large-scale datasets, some of which include real images and others consisting of synthetic images. In early detection methods, CNNs were used to make predictions about local segments, and then heuristics were used to combine those predictions. [19,67] Methods based on region proposal and bounding box regression weren't introduced until much later [36]. Component-level approaches have been proposed as of late, with a recent emphasis placed on the homogeneity and locality of texts. Numerous modern text recognizers all use the same methodology [37,53,52,59], which can be broken down into a combination of a few standard deep modules [3] if one were to interpret it that way. In the most recent generation of text recognition models, CNNs are used to encode the image into a feature space after the text instance image has been cropped. After that, a decoder is used so that characters can be extracted from the features.

Classification of the document type is the first step in automating document processing, and it is also the step that is considered to be the most important. The initial methods dealt with the problem as a matter of general image classification, which led to the testing of a number of different CNNs. In more recent times, with the introduction of BERT methods, which are based on a combination of CV and NLP, widespread proposals have

been made. The OCR-ed texts are then serialized into a token sequence after this step, and finally, they are fed into a language model (such as BERT) with some visual features if they are available. This completes the process. This is a tried-and-true method for extracting texts, and the vast majority of extraction strategies employ it. In spite of the fact that the idea is straightforward, the methods have shown significant improvements in performance and have become a primary industry trend in recent years. The range of practical applications is explored in detail in Document IE. For example, if a document parser is given a collection of raw receipt images, it is able to automate a significant portion of the process of receipt digitization.

This was previously something that required a large number of hours of manual labor from a human workforce in the conventional processing pipeline. The output of OCR is an input that is utilized by the overwhelming majority of modern models. Following the OCR, a series of subsequent processes, many of which are typically quite complicated, are utilized in order to convert the results of the OCR into the final parse. There have only been a few works that have attempted end-to-end parsing despite the fact that the industry has a need for it. This is despite the fact that there is a need for it. In recent years, a number of different works have been proposed with the objective of simplifying difficult parsing procedures. However, in order to extract text information, they continue to use an OCR that is independent of the other software. When performing visual quality assurance on documents, the goal is to provide answers to questions that are posed using screenshots of the documents. You will need to exercise reasoning that is based on the visible components of the image as well as general knowledge in order to correctly infer the answer to this problem. The majority of OCR systems that are currently considered to be state-of-the-art use a straightforward pipeline that begins with the application of BERT-like transformers and then moves on to OCR.

This method is currently considered to be the most effective. On the other hand, the procedures are intended to operate in a fashion that is, by definition, extractive in character. As a consequence of this, there are a few questions that need to be answered regarding the query for which the image that has been provided does not contain the answer. Methods that are based on generations are yet another strategy that has been suggested as a way to address the concerns.

IV. Findings and Conclusions

During the course of this research, we experimented with a variety of OCR software, including easy OCR, keras OCR, and others. Although such OCR-based approaches promise decent performance, they suffer from significant difficulties created by the OCR, such as (1) costly computational expenses and (2) performance degradation due to the OCR error propagation. Both of these drawbacks are caused by the OCR. As a result, we devised the Donut Document understanding transformer; this OCR-free end-to-end transformer concept is what we came up with. We are certain that this model will be an excellent fit for addressing any issues that we may have around OCR.

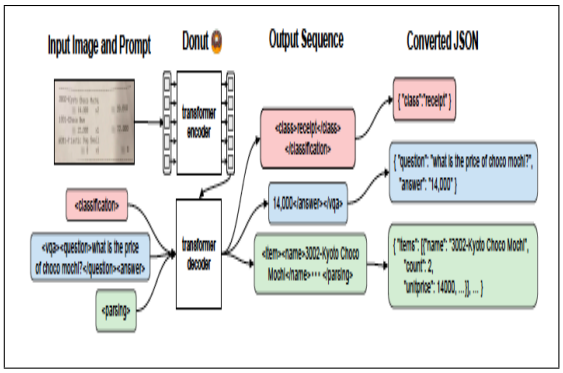

Fig. 3. Barnes-Hut domain decomposition

V. Future-Work

In the future, we would like to perfect the Donut with photographs of hotel menus and develop a more accurate method for separating text from images. After the generation of the text file, a jsonl-file containing the headings and prices will be obtained. This will be accomplished by creating a Json file and carefully calculating the bounding boxes of the image.According to the Json file, it is possible for us to develop a chat bot that will respond to the fundamental questions that are posed in the menu by the user.We anticipate a level of accuracy of 88 percent coming from this Donut model. This is high in comparison to the OCR's that are currently in effect.

References

[1] https://ieeexplore.ieee.org/abstract/document/9183326?casa$_{t}oken$=$CpOJBowDZXEAAAAA:ty45o24_{j}T9yBRCC1p6MvDXHx-c8tGReHMwBo1j8UWY0KKbrksjxRC8tSMdKcwD00zIxw3S0e3Uhttps://www.sciencedirect.com/science/article/abs/pii/S0925231213006309?via$

[2] https://dl.acm.org/doi/abs/10.1145/3503161.3548547?casa$_{t}oken$=$2BjH3kYeO2kAAAAA:https://link.springer.com/article/10.1007/s42001-021-00149-1$

[3] https://dl.acm.org/doi/abs/10.1145/3439726?casa$_{t}oken$=$pPPK-e1RrM8AAAAA:MK8DVNeImHf8x8fy_{q}-q5A9Z6OhI8IV-YbdlS-jU7CpzLdM0b5mau1PiJ-S_{Q}QoeMKU7AG2r9PFMKA:https://arxiv.org/abs/2204.03954$

[4] https://arxiv.org/abs/2111.15664

[5] https://arxiv.org/abs/1912.13318

[6] .https://ieeexplore.ieee.org/document/9342722

[7] https://ieeexplore.ieee.org/document/8748309

[8] .https://arxiv.org/abs/2010.11080

[9] https://ieeexplore.ieee.org/document/8953846

[10] https://ieeexplore.ieee.org/document/7333933

[11] https://aclanthology.org/N19-1423/

[12] https://towardsdatascience.com/ocr-free-document-understanding-with-donut-1acfbdf099be
[13] https://medium.com/capital-one-tech/learning-to-read-computer-vision-methods-for-extracting-text-from-images-2ffcdae11594