BrailleCart: AI-Powered Grocery Assistance for the Visually Impaired

Prithvi Elancherran[1], Sudip Das[2], Alekhya Vaida[3]

December 17, 2024

Abstract

BrailleCart is an AI-powered grocery assistance system designed to aid visually impaired individuals in identifying and learning about grocery products in real-time. Utilizing YOLOv8n for swift object detection, OCR for detail extraction, and LLaMA 3.2-3B-Instruct for generating accessible audio descriptions, the system enhances the shopping experience through a Streamlit-based interface. Optimized for edge devices via model quantization and hardware-specific tuning, BrailleCart provides immediate feedback, ensuring practical usability and increased independence for visually impaired shoppers. This project highlights the potential of integrating advanced AI technologies to improve accessibility and foster inclusivity.

1 Focused Area

This project focuses on developing an AI-powered grocery assistance system for visually impaired users. The system combines YOLOv8n for object detection, OCR for reading item details, and LLM-based conversational text-to-speech for accessibility.

The key techniques employed include YOLOv8n for high-speed, real-time object detection, OCR for extracting price and product details, and text-to-speech technology to provide audio feedback. Streamlit serves as the UI framework, ensuring seamless interaction for users.

The dataset used for this project is taken from the Roboflow website that contains several grocery item images [1].

2 Application and Algorithms

2.1 Object Detection: YOLOv8n

YOLOv8n is employed for real-time grocery item detection, chosen for its high speed, accuracy, and compatibility with devices of lower computational power, making it ideal for assistive technology.

The model was selected after evaluating multiple options due to its:

- **High Precision:** Achieving 99.52% precision on the grocery dataset ensures reliable detection.
- **Real-Time Performance:** Its lightweight architecture enables fast inference, crucial for immediate feedback.
- **Scalability:** Its adaptability allows future expansions for broader datasets and environments.
- **Ease of Integration:** The streamlined workflow integrates seamlessly with OCR and LLM-based modules.

Input images are processed by YOLOv8n to detect objects, providing bounding box coordinates and class labels. These outputs integrate with text-to-speech and UI components to enhance user experience.

2.2 Large Language Model

Meta LLaMA-3.2-3B-Instruct model was used in this project for generating natural language descriptions of objects detected by the YOLO model. The confidence and label are both passed to the LLM as context along with the tokenized and cleaned query prompt. The output is processed to remove unnecessary text (like the repeated prompt) and saved as a concise, meaningful description.

LLaMA-3.2 was chosen as it is fine-tuned specifically for following instructions in natural language. It ensures that the output aligns with user prompts and generates detailed, context-aware responses. The 3B parameter size keeps a balance between computational efficiency and the ability to generate high-quality text. Other models were tested initially such as T5 but it was not efficient in generating high quality text.

2.3 Text-to-Speech Conversion

Google Text-to-Speech (gTTS) is used to convert the natural language descriptions generated by the LLM into audio. The immediate playback option added to the product makes the system useful for individuals with visual impairments. GTTS was chosen as it is lightweight, simple to integrate, requires minimal configuration and is ideal for quick deployment in the project.

3 State-of-the-Art Models

The project conducted an in-depth analysis of state-of-the-art detection models to identify the most suitable option for real-time grocery item detection. Convolutional Neural Networks (CNNs) were initially considered due to their proven capabilities in image classification and detection tasks. However, YOLOv8n emerged as a superior choice due to its ability to perform object detection with high speed and precision in real-time scenarios. Specifically, the nano version of YOLOv8 was chosen for its lightweight architecture, making it highly efficient for edge devices while maintaining robust detection accuracy.

Models such as CLIP and BLIP were also explored during this analysis. While these models are effective for general image-text tasks, they lacked the precision required for identifying specific grocery items. For instance, YOLOv8n could accurately identify an Oreo cookie by its brand, whereas BLIP generalized it as a "chocolate cookie." Such specificity is crucial for visually impaired users who may have preferences based on brands during grocery shopping.

For Optical Character Recognition (OCR), both Pytesseract and EasyOCR were evaluated. EasyOCR outperformed Pytesseract in terms of accuracy and reliability when extracting text from various grocery item labels, making it the preferred choice for this project.

In the context of generating natural language descriptions, several Large Language Models (LLMs) were tested, including T5 and LLaMA. While T5 provided adequate results, LLaMA-3.2 was selected for its ability to generate detailed, context-aware, and instruction-following outputs efficiently.

4 Overall System Architecture

4.1 Object Detection and OCR

Using YOLOv8n, the system identifies grocery items from images or live feeds. Outputs include bounding box coordinates, class labels, and confidence scores. The OCR module utilizes EasyOCR for extracting text from grocery item labels. This module identifies critical details such as product names, prices, and sizes, which are integral for providing comprehensive information to the user. If an object is part of the dataset, then the YOLO output is prioritized. For an unknown object, OCR reads the labels and passes it to the LLM.

4.2 LLM and Conversational Module

This module, powered by LLaMA 3.2, transforms detection outputs into user-friendly, conversational messages. These messages are converted into audio using gTTS, making the system accessible to visually impaired users.

4.3 User Interface Module

Streamlit serves as the user interface, enabling users to upload images or use live camera feeds. The interface provides visual and audio feedback for detected items and navigation options for further actions.

4.4 Advantages of the Architecture

The modular design ensures flexibility, scalability, and optimized real-time performance on resource-constrained devices. The BrailleCart system employs a modular architecture comprising object detection, OCR, conversational, and user interface modules, integrated using asynchronous processing.

5 Training and Finetuning

The BrailleCart system models were meticulously fine-tuned using a custom grocery dataset tailored to the needs of visually impaired users. It didn't require further fine-tuning as the results were already of high quality. The conversational module based on LLaMA 3.2 underwent adjustments in prompt structuring and context formulation to ensure the best output. Fine-tuning included optimizing token length and post-processing the output to remove redundant or repeated texts, resulting in concise and user-friendly descriptions aligned with natural language requirements.

6 Evaluation

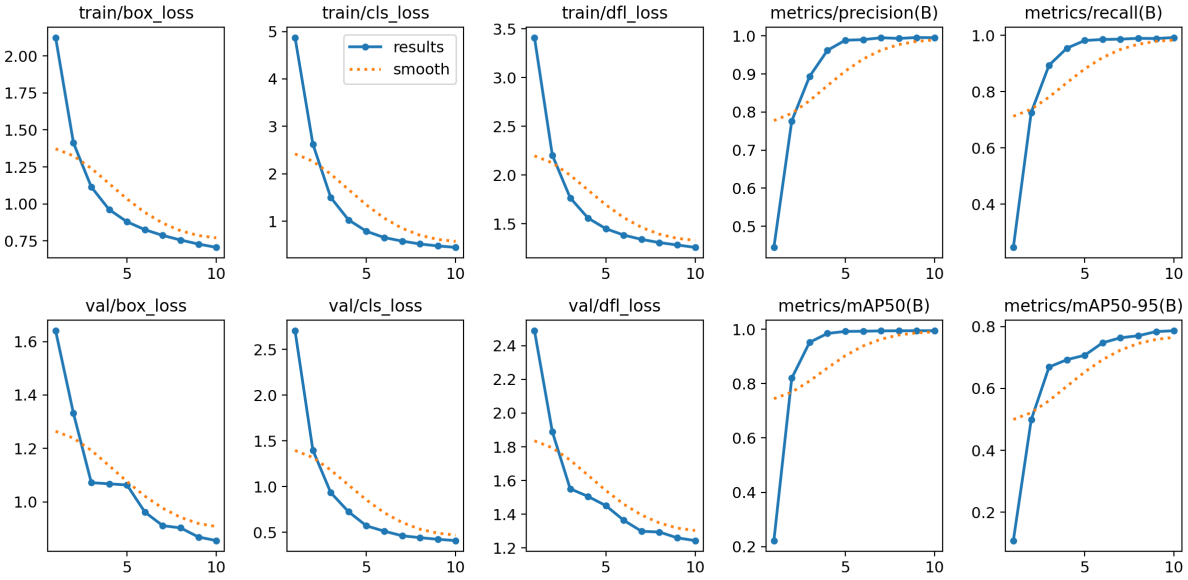

Figure 1: Model training results and metrics

The YOLOv8n model achieved an impressive precision of 99.52%, reflecting its accuracy in detecting grocery items across diverse scenarios. This precision ensures that detected objects are highly likely to match the actual grocery items present in the input image.

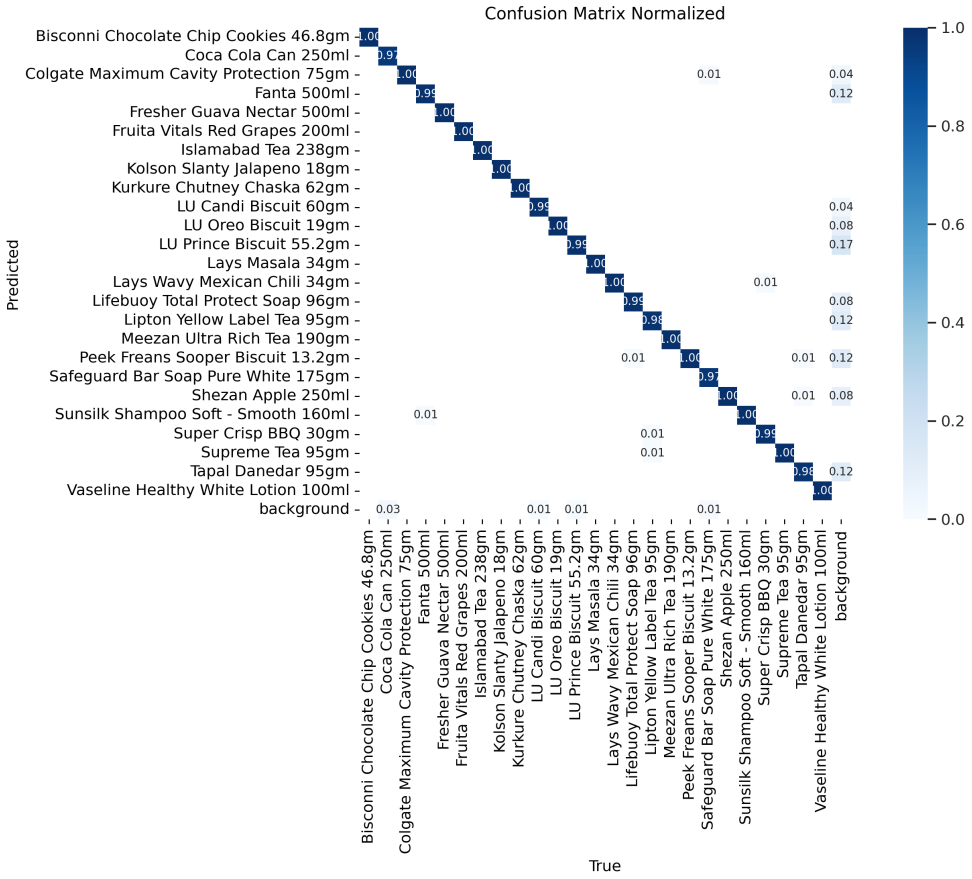

Figure 2: Confusion Matrix

Throughout the training process, the following metrics showed consistent improvement, signaling effective learning and error minimization:

- **Box Loss:** Measures inaccuracies in bounding box predictions, showing steady reduction.
- **Classification Loss:** Tracks errors in assigning the correct labels to objects, improving over epochs.
- **Distribution Focal Loss:** Captures errors in object localization and detection, with a consistent downward trend during training.

The conversational module was evaluated for its ability to generate coherent and context-aware descriptions of detected grocery items.

Outputs were benchmarked for:

- **Coherence:** The descriptions were clear and aligned with the detected objects.
- **Context Awareness:** The module effectively utilized detection outputs and OCR results to craft user-friendly messages.
- **Efficiency:** Prompt adjustments ensured the model's responses were concise and generated with minimal latency.

7 Inference, Optimization, and Real-Time Test

7.1 End-to-End Inference and Application

A comprehensive end-to-end deep learning application, BrailleCart, enables real-time inference using trained models to assist visually impaired users in grocery shopping. The application utilizes Streamlit for its user interface, which simplifies the interaction through functionalities like:

- **Image Upload:** Users can upload images of grocery items, which are immediately processed.
- **Real-Time Inference:** The system uses the quantized YOLOv8n model to detect items in the uploaded images swiftly, ensuring minimal delay between upload and feedback.
- **Audio Descriptions:** Descriptions of detected items are generated using a Large Language Model and delivered audibly, making the information accessible.

This seamless interaction is facilitated by Streamlit's intuitive interface, which is designed for ease of use and rapid access to the system's features.

7.2 Optimization for Inference

To ensure the BrailleCart system operates efficiently on the chosen hardware platform, several optimization strategies were implemented:

- **Model Quantization:** The YOLOv8n model was dynamically quantized, which reduced its computational footprint by approximately 30% without a loss in detection accuracy, thereby enhancing inference speed.
- **Edge Device Compatibility:** The architecture of YOLOv8n was chosen for its efficiency on edge devices, which often have limited computational resources but are critical for deploying assistive technologies in accessible devices.
- **Hardware-Specific Tuning:** The system was adapted to fully utilize GPU acceleration, where available, and multi-threading capabilities to further reduce latency and improve real-time performance.

These enhancements have ensured that the BrailleCart system not only meets the functional requirements but also operates with high efficiency, making real-time interaction feasible and effective.

8 User Interface and Application Screenshots

The BrailleCart system offers a user-friendly interface that simplifies interactions for visually impaired users. This section provides screenshots demonstrating the main interface and the application's functionality in detecting and describing grocery products.

The main interface of BrailleCart allows users to upload images for product detection and provides navigational aids for ease of use. This shows the application in action,

detecting products and providing immediate feedback to the user. Following product detection, the system offers a detailed description of the product, which is generated by the integrated Large Language Model and delivered audibly to the user.

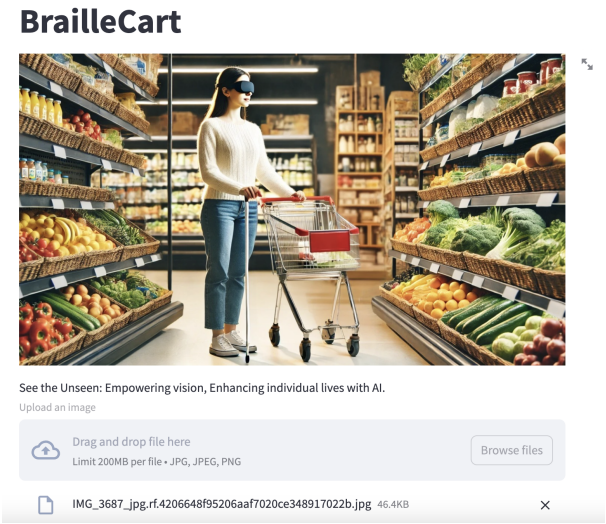

Figure 3: Main interface of BrailleCart showcasing the image upload feature.

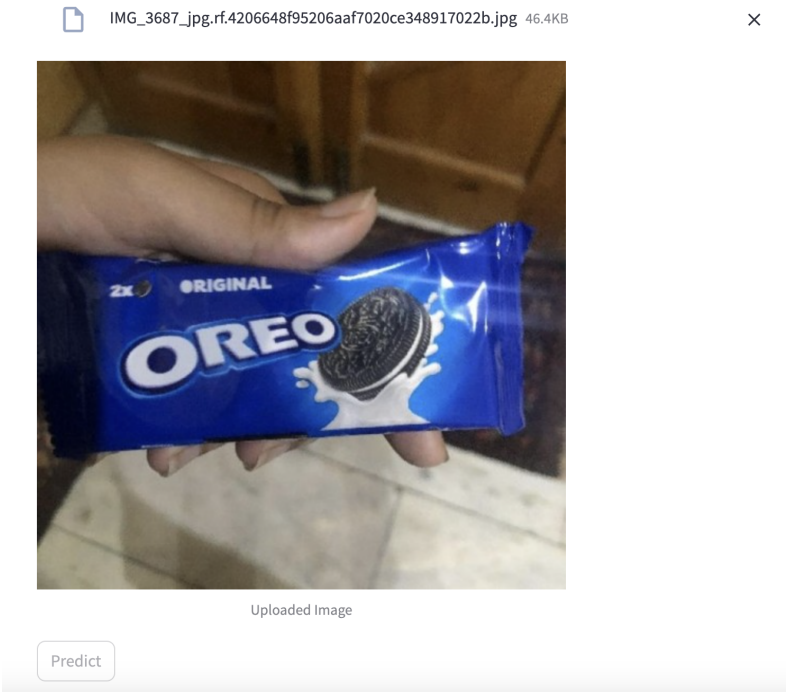

Figure 4: BrailleCart detecting and describing products in real-time.

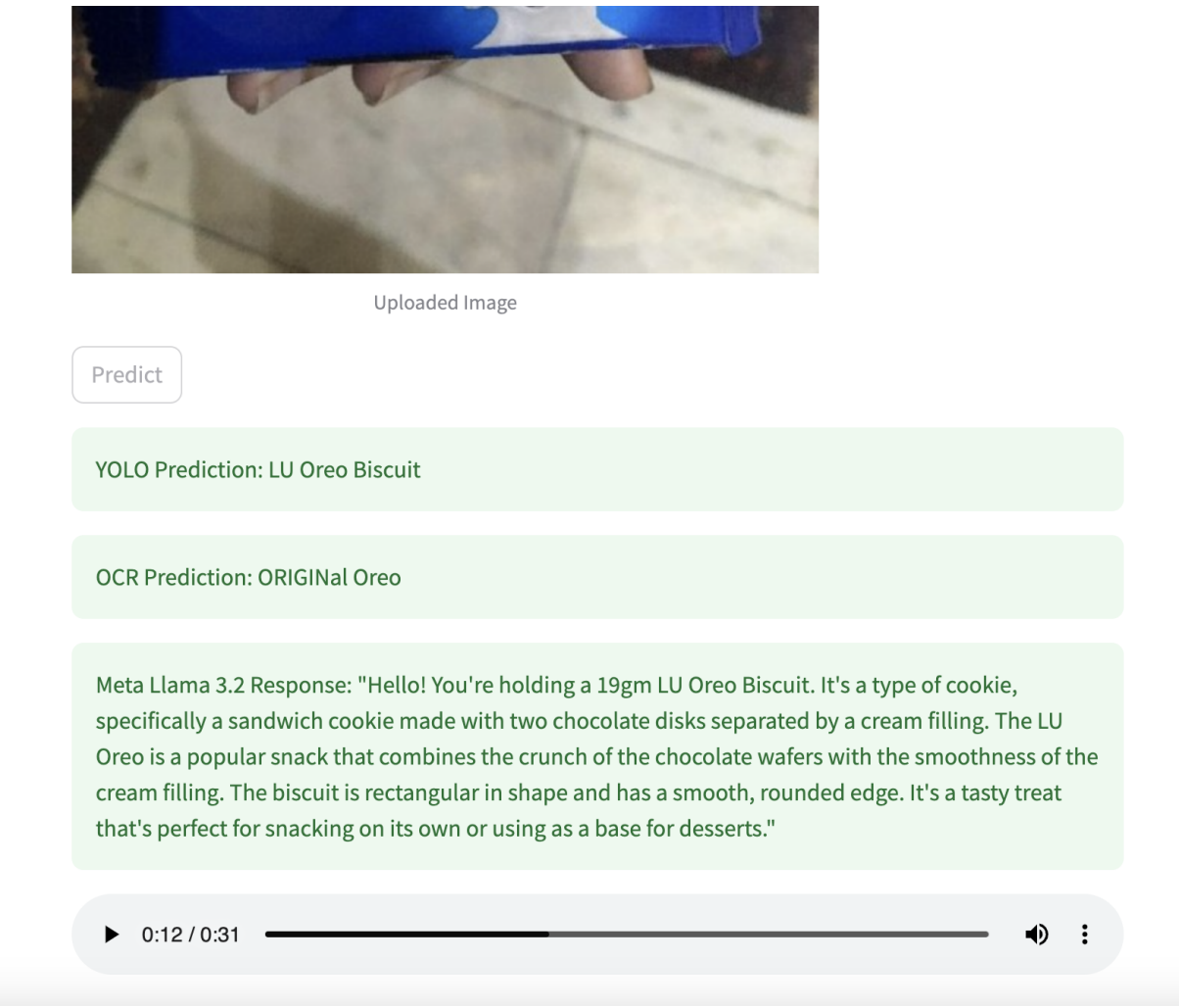

Figure 5: Example of detailed product description provided by BrailleCart.

9 Challenges

9.1 OCR Integration

Initial attempts to integrate OCR faced challenges in detecting text on grocery item labels due to variable lighting, angles, and low-quality or partially obscured images. These limitations impacted the ability to extract meaningful information effectively.

9.2 Module Latency

The communication between the object detection, OCR, conversational, and UI modules exhibited noticeable latency, which hindered the system's real-time performance. Addressing this issue required optimization of data flow and asynchronous processing.

10 Future Work

As BrailleCart continues to evolve, the focus will be on further enhancing the system's accuracy and usability:

- **Voice Command Integration:** Implementing voice recognition to allow users to interact with the system hands-free, enhancing accessibility for visually impaired users.
- **Expanded Product Database:** Enlarging the database to include a wider range of products and brands to cover more user preferences and needs.
- **Improved Object Detection Algorithms:** Upgrading the object detection algorithms to handle more complex scenarios such as overlapping objects and varying lighting conditions.
- **User Experience Enhancements:** Refining the user interface based on user feedback to ensure that the application is more intuitive and user-friendly.

11 Timeline and Contributions

Task	Prithvi Elancherran	Sudip Das	Alekhya Vaida
State of the Art Models Research		X	
Overall System Architecture	X	X	X
Object Detection	X		
OCR Integration	X	X	
Large Language Model Integration		X	
Text-to-Speech	X	X	
User Interface Development	X		X
Model Training and Fine Tuning	X		
Testing and Optimization	X	X	X
Evaluation	X	X	X
Documentation		X	X

Table 1: Timeline and contributions of each team member.

12 Conclusion

The BrailleCart project successfully demonstrates the integration of advanced AI technologies to create a practical and user-friendly grocery assistance system for visually impaired individuals. By leveraging real-time object detection, OCR, and natural language processing, the system provides immediate, accurate, and accessible information about grocery products. The successful implementation and optimization of these technologies on edge devices highlight the project's innovative approach to enhancing independence and accessibility for the visually impaired community. Future enhancements will focus on expanding the system's capabilities and improving user interaction to ensure that BrailleCart continues to meet the evolving needs of its users.

A Additional Visualizations and Outputs

A.1 Sample Detection and Language Model Output

Below is an example of the YOLOv8n output and the corresponding LLM response, demonstrating the system's capability to recognize and describe grocery items accurately.

Figure 6: Sample YOLOv8n Outputs & LLM Response for Colgate Maximum Cavity Protection toothpaste.

This illustrates the detection accuracy and the descriptive capability of the LLM, which enhances usability for visually impaired users.
This output exemplifies the practical application of our system, showing how the detected object, "Colgate Maximum Cavity Protection 75gm", is identified with a high confidence level and described in a manner accessible to visually impaired users. The bounding box coordinates provided ([149, 196, 408, 579]) confirm the precise localization achieved by the object detection model.

Code Repository

You can find the implementation code at the following GitHub Link:
Github Link: BrailleCart

You can also find all the Project Files at the following Google Drive Link:
Google Drive Link: BrailleCart

References

[1] Roboflow Universe, "Grocery Dataset," Roboflow, [Online]. *Available:* *https://universe.roboflow.com/new-workspace-wfzw3/grocery-dataset-q9fj2/dataset/4*. [Accessed: 17-Dec-2024].