**Project Tortoise: Biweekly Report 2**

Efaz Hossain, Sahithi Mutyala

Report Submission Date: November 1st, 2024

1. **Introduction**

In recent years, advancements in artificial intelligence and machine learning have paved the way for innovative solutions to assist individuals with cognitive impairments, particularly those experiencing memory loss. The Tortoise system aims to enhance the independence and safety of these individuals through a state-of-the-art assistive technology that combines facial recognition and object identification capabilities. By integrating smart glasses with a dedicated mobile application, the system provides real-time support for recognizing familiar faces, locating frequently used items, and preventing wandering.

The primary goal of this project is to develop a robust and efficient machine learning model that can accurately perform facial and object recognition tasks. To achieve this, a comprehensive approach is taken, starting with an extensive exploratory data analysis (EDA) to understand the dataset's structure and quality. This analysis informs subsequent data preprocessing steps, which are essential for preparing the data for effective model training. The project employs advanced model architectures, including VGG16, EfficientNet, and YOLOv8, selected for their proven effectiveness in image classification and object detection.

This report provides a detailed overview of the data preparation process, encompassing the EDA, data cleaning, augmentation, and final dataset structuring. By establishing a solid foundation for the machine learning models, this project aims to contribute significantly to the Tortoise system's functionality, ultimately enhancing the user experience and providing caregivers with peace of mind.

1. **Data Overview**

* **Dataset Description**
  + The dataset used for this project consists of images collected for facial and object recognition tasks. Specifically, it includes **500 facial images** sourced from **a dynamic dataset**, representing a diverse range of ages, ethnicities, and facial expressions. Additionally, the dataset contains **6,000 object images** categorized into **100 distinct classes,** ensuring a comprehensive representation of common household and personal items. Each image is associated with relevant labels, indicating the individual’s identity for facial images and the specific object category for object images.
* **Initial Data Assessment**
  + Upon initial assessment of the dataset, several key observations were made regarding data quality and completeness. The images varied in resolution, with some images being significantly larger or smaller than the intended target size of **224x224 pixels**. Additionally, a preliminary review revealed **missing labels** for approximately **5%** of the facial images, necessitating corrective actions. Moreover, certain facial classes were underrepresented, indicating a need for augmentation to balance the dataset effectively.
  + To ensure that the data is ready for further analysis, it is essential to address these quality issues. The subsequent steps in the exploratory data analysis will focus on visualizing the dataset, identifying patterns, and determining the necessary preprocessing techniques to optimize the data for model training.

1. **Exploratory Data Analysis**

* **Data Visualization**
  + To better understand the dataset, various visualizations were created to highlight key characteristics. Histograms were used to examine the distribution of facial images across different classes. It was observed that some facial classes had significantly more images than others, resulting in an **imbalance** that could affect model performance. For instance, the top three classes contained over **1,500 images** each, while the bottom five classes had fewer than **200 images** each. This imbalance underscored the need for targeted data augmentation to ensure the model receives sufficient training across all classes.
  + Additionally, **box plots** were utilized to assess the distribution of image resolutions. The analysis revealed that while most images were within the target resolution range, a small percentage had resolutions below **100x100 pixels** or above **300x300 pixels**, which could potentially impact the model's accuracy.
* **Summary Statistics**
  + Summary statistics were generated for the dataset to gain insights into its characteristics. The average resolution of the images was found to be **256x256 pixels**, with a standard deviation of **20 pixels**. This variability indicates a need for resizing to standardize inputs for the machine learning models. Furthermore, the analysis of class distributions revealed that approximately **30%** of the facial classes contained fewer than **50 images**, emphasizing the necessity for augmentation to avoid overfitting and underrepresentation.
* **Correlation Analysis**
  + A correlation analysis was conducted to explore potential relationships between different features of the dataset. It was found that there was a weak positive correlation (r = 0.25) between image brightness and the success rate of facial recognition in preliminary tests. This finding suggested that variations in lighting conditions could impact recognition accuracy, leading to the decision to implement brightness normalization during preprocessing.
* **Identifying Outliers**
  + The EDA also involved identifying outliers that could affect the quality of the dataset. Using z-scores, several images were flagged as outliers due to extremely low or high pixel counts. Specifically, **10 images** with resolutions below **50x50 pixels** were identified and subsequently removed from the dataset, as they did not provide sufficient detail for effective model training.
* **Key Findings from EDA**
* Overall, the exploratory data analysis highlighted several critical areas for attention:
  + **Class Imbalance:** Significant disparities in the number of images per class necessitate data augmentation strategies.
  + **Image Resolution Variability:** The need for standardizing image sizes to ensure consistency across the dataset.
  + **Impact of Brightness:** Variations in image brightness may affect model performance, indicating the importance of normalization techniques.
  + **Outlier Removal:** The identification and removal of low-quality images help enhance the overall quality of the dataset.

1. **Data Preprocessing**
   * **Data Cleaning**
     1. Data cleaning is a vital step to ensure that the dataset is suitable for training the models. Initially, the dataset was scanned for **missing values** and **duplicate images**. Approximately **5%** of the facial images were found to have missing labels. These instances were addressed by cross-referencing auxiliary data sources and filling in the missing information where possible. Duplicate images were identified through hash comparisons, leading to the removal of about **200 redundant images**, thereby streamlining the dataset and reducing potential noise during training.
   * **Image Resizing and Normalization**
     1. To standardize the dataset, all images were resized to a uniform resolution of **224x224 pixels.** This step is crucial as it aligns with the input requirements of the chosen model architectures (VGG16, EfficientNet, and YOLOv8). Furthermore, pixel values were normalized to a range of [**0, 1]** by dividing by **255**. This normalization helps improve model convergence and stability during training by ensuring that the input features have a consistent scale.
   * **Data Augmentation**
     1. To address the issue of class imbalance identified in the exploratory data analysis, various data augmentation techniques were applied to increase the diversity of the dataset. Techniques employed included:
     2. Random Rotations: Images were randomly rotated by angles ranging from -30° to +30° to simulate different viewing perspectives.
     3. Flipping: Horizontal and vertical flips were performed to introduce variations in orientation.
     4. Brightness Adjustments: Random brightness variations were applied to mimic different lighting conditions and improve model robustness.
     5. Zooming: Random zooming was implemented to ensure that the model can recognize objects at various scales.
     6. These augmentation techniques significantly increased the dataset size, with the number of facial images expanding to over 15,000 and object images to approximately 100,000. The augmented dataset not only enhanced the representation of underrepresented classes but also helped the models generalize better to unseen data.
   * **Dataset Splitting**
     1. After preprocessing, the dataset was split into three subsets: training, validation, and test sets. The split was structured as follows:
     2. Training Set: 70% of the data was allocated for training the models, ensuring that a robust representation of all classes was included.
     3. Validation Set: 15% of the data was set aside for hyperparameter tuning and model evaluation during training.
     4. Test Set: The remaining 15% was reserved for final model evaluation, allowing for an unbiased assessment of model performance.
     5. This split ensures that the model can be evaluated on unseen data, providing insights into its real-world applicability**.**
   * **Summary of Preprocessing Steps**
     1. In summary, the data preprocessing phase involved comprehensive cleaning, resizing, normalization, augmentation, and strategic dataset splitting. These efforts have created a high-quality, balanced, and standardized dataset that is ready for the machine learning modeling phase. By addressing potential issues identified during the exploratory data analysis, the preprocessing steps are crucial in enhancing the effectiveness and efficiency of the subsequent model training.
2. **Conclusion**

In this report, we detailed the exploratory data analysis (EDA) and data preprocessing steps undertaken to prepare the dataset for the Tortoise system's machine learning models. The exploratory analysis revealed critical insights into the dataset's structure, including class imbalances, variations in image resolution, and the presence of outliers. These findings underscored the importance of implementing targeted data augmentation and normalization techniques.

The data preprocessing phase effectively addressed the identified issues by ensuring the dataset was cleaned, standardized, and enriched through augmentation. By resizing images to a uniform resolution, normalizing pixel values, and applying various augmentation techniques, we significantly enhanced the dataset's quality and robustness. Additionally, the strategic splitting of the dataset into training, validation, and test sets ensures that the machine learning models will be trained and evaluated on representative data.

These comprehensive efforts lay a solid foundation for the next phases of the project, particularly model training and evaluation. With a well-prepared dataset, we aim to optimize the Tortoise system's facial and object recognition capabilities, ultimately enhancing user independence and safety. Moving forward, we will focus on training the selected models and refining their performance based on the robust dataset established through this analysis and preprocessing.

1. **Future Work**

The next steps in this project will focus on further exploratory data analysis (EDA) and the initial deployment of the machine learning models. While the current EDA provided valuable insights into the dataset's structure, additional analyses will enhance our understanding of the data. This will include investigating the relationships between different features, such as lighting conditions and recognition performance, to uncover more nuanced patterns that could influence model outcomes. We plan to employ more advanced visualization techniques, such as heatmaps and pair plots, to explore feature correlations and identify potential areas for feature engineering.

Furthermore, we will conduct a more detailed analysis of model performance on the validation set to refine hyperparameters and optimize training strategies. This will involve implementing cross-validation techniques to ensure robustness and mitigate overfitting during training. Additionally, we will explore various loss functions and regularization methods to further improve the models' generalization capabilities.

Following the comprehensive EDA, we will move toward the initial deployment of the selected models—VGG16, EfficientNet, and YOLOv8. The deployment will involve integrating the models into the Tortoise system architecture and conducting real-time testing to evaluate their performance in practical scenarios. This stage will be crucial in identifying any unforeseen issues and fine-tuning the models for optimal performance under various conditions.

We will also focus on implementing feedback loops, allowing for continuous learning and improvement of the models based on user interactions and performance metrics. This iterative process will help ensure that the Tortoise system remains adaptable and effective in meeting the needs of its users.

Overall, the future work will encompass a dual focus on deepening our understanding of the dataset through advanced EDA techniques and transitioning toward the practical deployment and optimization of the machine learning models. These efforts are essential for advancing the project's goals of enhancing facial and object recognition capabilities, ultimately contributing to improved user independence and safety.