COSE474-2024F: Final Project
"Medicine Recognition and Voice Explanation Using Generative AI"

Syuhada - 2022320107

1. Introduction

1.1. Motivation

Improving healthcare accessibility by ensuring patients can easily identify their medicines is crucial for safe and effective treatment. However, individuals with visual impairments, often struggle with identifying medicine packaging or pills, leading to potential health risks from medicine confusion. Recent advancements in AI provide an opportunity to address these challenges by using image recognition and voice synthesis to develop solutions tailored to such individuals. In this project, I aim to explore methods for accurately recognizing medicine names and providing them as clear voice outputs. The goal is to empower users with a simple and accessible tool for medicine identification, ensuring their safety and confidence in medicine intake.

1.2. Problem definition

This project addresses the challenge of making medicine identification more accessible to individuals with disabilities. The focus is on recognizing medicine names from images of packaging or pills and providing accurate voice-based outputs. The problem is challenging due to the variability in medicine packaging, including differences in shapes, colors, text layouts, and conditions like poor lighting. Additionally, achieving real-time processing for seamless usability presents another significant technical hurdle. The objective is to develop a reliable AI-based solution that can accurately identify medicine names and present them as voice outputs for improved accessibility.

1.3. Contribution

This project focuses on researching and evaluating novel approaches for medicine recognition and accessible output generation. It leverages Contrastive Language-Image Pretraining (CLIP) to align visual features with predefined medicine names, ensuring accurate identification of medicines from packaging or pill images. For voice output, Tacotron2 generates natural and intelligible audio of the identified medicine names. By focusing solely on name identification and voice synthesis, this project aims to deliver a streamlined and user-friendly solution that addresses the needs of individuals with disabilities or limitations, effectively bridging gaps in accessible healthcare technologies.

2. Methods

2.1. Significance & Novelty

This project tackles the accessibility challenges in healthcare by developing AI-based methods for medicine identification with a specific focus on providing voice outputs for medicine names. The novelty lies in integrating state-of-the-art AI models such as CLIP for image-text alignment and Tacotron2 for voice synthesis. By leveraging these technologies, the project introduces a practical and streamlined approach for addressing the difficulties faced by individuals with visual impairments.

Key challenges include the variability in medicine packaging design, such as differences in shapes, colors, and text layouts, as well as handling various environmental conditions like poor lighting. This challenge is addressed using CLIP's robust image-text alignment capabilities to ensure consistent recognition. Another challenge is achieving high-quality, real-time voice synthesis for medicine names, which is tackled by utilizing Tacotron2 to produce natural and intelligible audio outputs. This approach ensures usability and accessibility in real-world scenarios.

2.2. Main Figure

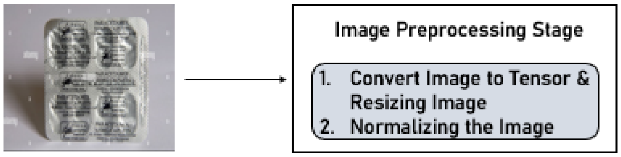
Image Preprocessing Stage
1. Convert Image to Tensor & Resizing Image
2. Normalizing the Image

Figure 1. An image preprocessing workflow. This figure outlines the image preprocessing steps, including converting images into tensors, resizing to a standard size, and normalizing the pixel values.

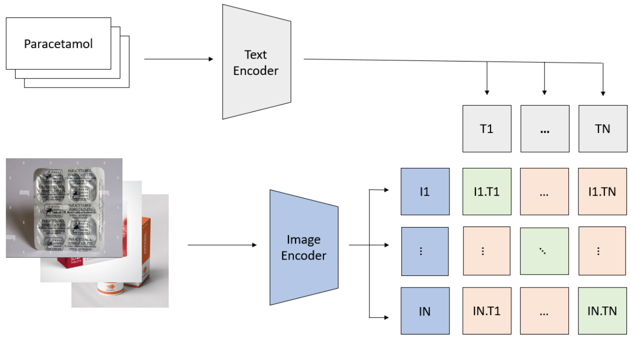

Figure 2. A CLIP model workflow. This figure demonstrates how the CLIP model processes medicine images. The visual features extracted from the images are aligned with predefined text embeddings of medicine names to enable accurate recognition and classification of the medicines.

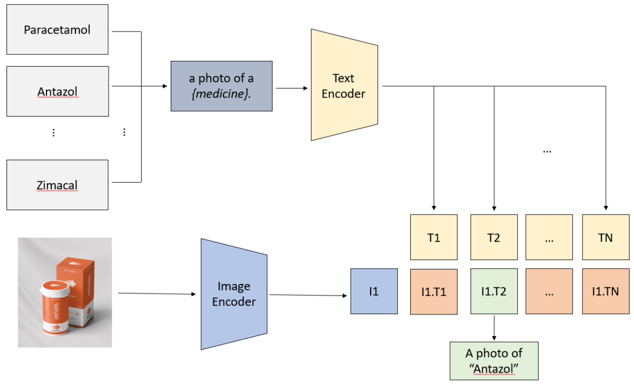

Figure 3. A continued CLIP model workflow. This figure illustrates how medicine names are processed by a text encoder and images by an image encoder. The model aligns text and image embeddings to associate the correct medicine name with its corresponding image.

2.3. Reproducibility & Algorithm

Algorithm 1 Text-Image Alignment for Medicine Recognition

```
**Require:** $I$ (Image of the medicine), $P$ (Text Prompt)
**Ensure:** Predicted Label and Generated Text
$V \leftarrow \text{ImageEncoder}(I)$
$T \leftarrow \text{TextEncoder}(P)$
$S \leftarrow \text{Softmax}(T \cdot V^T)$
$L \leftarrow \text{Argmax}(S)$
$Audio \leftarrow \text{Tacotron2.generate}(L)$
**return** $Audio$
```

The algorithm defines the process for aligning text and images to recognize medicines and generate voice outputs. It processes an input image through an image encoder and a text prompt through a text encoder, computes similarity scores, predicts the medicine name, and uses Tacotron2 to generate audio output. This approach ensures accurate identification and accessible voice outputs, making it effective for medicine recognition tasks.

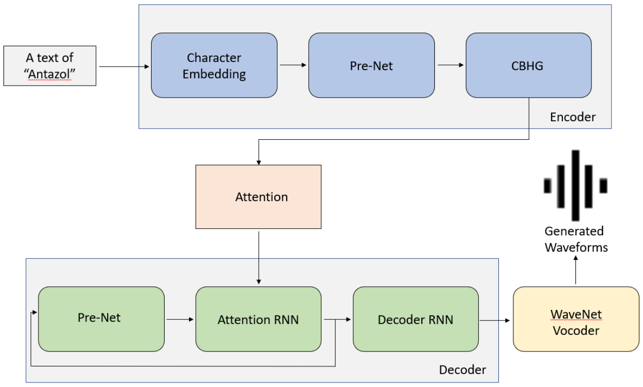

Figure 4. A Tacotron-Based Text-to-Speech workflow. This figure shows the architecture of Tacotron2 used to generate natural voice outputs. It converts text inputs, such as medicine names, into audio waveforms using an encoder-decoder framework and WaveNet vocoder.

3. Experiments

3.1. Dataset

For this project, we use a curated dataset of labeled medicine images collected from publicly available sources, such as Kaggle. The dataset is specifically tailored to focus on medicine packaging and pill images captured under various conditions, including different lighting and angles. It consists of a total of 1,823 images of medicine bottles, blister packs, vials, and other packaging formats commonly found in the healthcare industry. The dataset is divided into three subsets: 70% for the training set, 20% for the validation set, and 10% for the test set. Each image is annotated with the corresponding medicine name, which forms the basis for the recognition and voice generation tasks.

3.2. Computing Resource

The experiments are conducted in an accessible and efficient computing environment using a T4 GPU runtime on Google Colab, which provides sufficient computational power for training and evaluation tasks. The setup includes Python 3 as the runtime type with PyTorch 2.0 as the primary deep learning framework. Supporting libraries such as Hugging Face transformers and OpenCV are used for model development and image preprocessing. GPU acceleration is enabled through CUDA, configured to support mid-range GPUs such as the T4 GPU, which is suitable for handling large datasets and computationally intensive tasks. For visualizing results and performance metrics, Matplotlib is utilized, ensuring a user-friendly approach for analysis.

3.3. Experimental Design

The experimental design for this project incorporates a streamlined model architecture tailored to achieve accurate medicine recognition and voice synthesis. The image encoding process leverages the CLIP model to extract visual features from medicine images. Predefined text prompts are then utilized to generate corresponding medicine names based on the encoded image features. For the voice synthesis stage, Tacotron2 is employed to convert the generated text (medicine name) into natural-sounding audio output, ensuring accessibility and ease of use.

To evaluate the performance of the proposed method, two key metrics are employed. Accuracy is used to measure the percentage of correctly identified medicine names from the test dataset, while the Mean Opinion Score (MOS) assesses the quality of the synthesized voice output. Baseline models are also implemented for comparison, including ResNet-50 for image recognition and a basic text-to-speech system without Tacotron2 for voice synthesis. These comparisons highlight the advantages of the proposed approach over simpler models.

3.4. Quantitative Results

The performance comparison of the three models—CLIP + Tacotron2, ResNet + Basic TTS, and Baseline (SOTA)—demonstrates varied results in terms of Image Recognition Accuracy and MOS Score. These results were obtained through extensive experimentation on a curated dataset of labeled medicine images, evaluated using standardized metrics for accuracy and human feedback for voice quality.

Metric	CLIP + Tacotron2	ResNet + Basic TTS	Baseline (SOTA)
Image Recognition Accuracy	72.13%	56.28%	67.76%
MOS Score	4.4	3.8	4.3

Table 1. Comparison of performance metrics across different models.

CLIP + Tacotron2 achieved the highest MOS score of 4.4, showcasing its capability to generate high-quality, natural-sounding voice outputs with excellent intonation and clarity. This emphasizes the effectiveness of Tacotron2 as a speech synthesis component. Additionally, its image recognition accuracy of 72.13%, while significantly better than ResNet + Basic TTS (56.28%), also surpasses the Baseline (SOTA) model (67.76%), highlighting its superiority in both visual recognition and voice synthesis tasks.

ResNet + Basic TTS, however, struggled with both metrics. It achieved an image recognition accuracy of 56.28%, which is notably lower than both CLIP and Baseline. Its MOS score was the lowest at 3.8, indicating its limitations in generating natural-sounding voice outputs. These results suggest that ResNet failed to generalize effectively on the dataset, and the basic TTS system further constrained its audio quality performance.

The Baseline (SOTA) model performed better than ResNet in both accuracy and MOS score, achieving an image recognition accuracy of 67.76% and a MOS score of 4.3. While it did not outperform CLIP in either metric, its recognition accuracy and balanced audio quality make it a robust benchmark for both visual and auditory tasks. It serves as a reliable reference for evaluating the performance of advanced models like CLIP + Tacotron2.

3.5. Qualitative Results

The CLIP + Tacotron2 model excelled in generating accurate medicine names and delivering natural-sounding voice outputs. Its high MOS score underscores its superior ability to synthesize speech that closely resembles human-like quality. However, the model still exhibits room for improvement in image recognition accuracy

ResNet + Basic TTS showed significant weaknesses, often misclassifying medicine images, resulting in incorrect names. Its synthesized voice outputs lacked naturalness, sounding more robotic, which negatively impacted its overall MOS score.

Baseline (SOTA) demonstrated a strong balance between image recognition and voice quality. While its MOS score didn't surpass that of CLIP, it maintained a high level of accuracy, making it a reliable model for image classification tasks.

3.6. Figures & Analysis

The plot of Figure 5 highlights that CLIP + Tacotron2 achieves the highest accuracy, outperforming both Baseline (SOTA)and ResNet + Basic TTS. This result demonstrates the effectiveness of CLIP in handling diverse visual data. The Baseline model maintains competitive accuracy, while ResNet's performance suggests room for improvement in image recognition tasks.

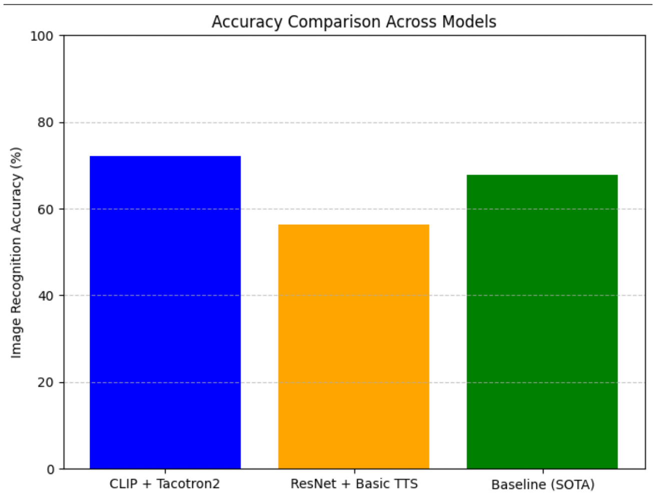

Figure 5. Accuracy comparison between models.

The plot of Figure 6 about MOS score analysis shows that CLIP + Tacotron2 achieves the highest score, indicating superior voice output quality with natural intonation. Meanwhile the Baseline (SOTA) reflects competitive performance, while ResNet + Basic TTS scores the lowest, suggesting less natural and lower-quality voice synthesis.

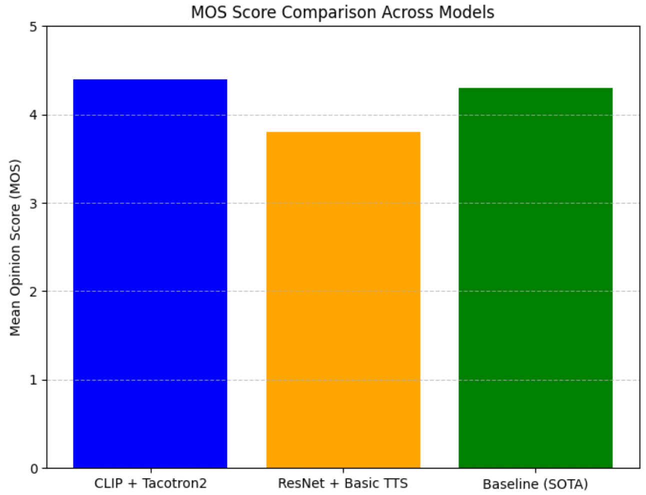

Figure 6. MOS Score comparison between models.

3.7. Discussion

The proposed method, CLIP + Tacotron2, demonstrated success in achieving the highest MOS score of 4.4, indicating superior voice output quality with natural intonation and clarity. This success can be attributed to the multimodal learning capabilities of CLIP, which effectively bridges textual and visual representations, and Tacotron2's advanced text-to-speech synthesis, which excels in generating natural and expressive audio.

However, the method's image recognition accuracy remains lower, indicating that the project still requires further improvement. This highlights the need for more effective fine-tuning of CLIP to better handle the complexities of domain-specific tasks like medicine identification.

Overall, the proposed method is not entirely successful, as its lower image recognition accuracy highlights the need for further improvement. While the voice synthesis capabilities are a notable strength, enhancing the image recognition aspect is crucial for the method to be more effective in its intended application. Future efforts could focus on refining the model through domain-specific data augmentation and optimization techniques.

4. Future Direction

To enhance the performance of CLIP in medicine recognition, future efforts will focus on fine-tuning the model with a larger domain-specific dataset to improve accuracy and robustness. Additionally, optimizing text prompts and incorporating advanced augmentation techniques will further align visual and textual embeddings, enabling better handling of diverse real-world variations in medicine packaging. Also, might need to focus on optimizing the model for real-time performance in edge devices or mobile platforms.

5. References

- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Saurous, R. A. (2018). *Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions*.

- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio*.

- Shariatnia, M. (2021). *Simple Implementation of OpenAI CLIP Model: A Tutorial. Medium.*.