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

Syuhada - 2022320107

1. Introduction

The motivation behind this project is to address the accessibility challenges faced by individuals with disabilities—particularly those who are visually impaired or have cognitive impairments—in accurately identifying medications. Misidentifying or misunderstanding medicine can lead to serious health risks. Currently, medicine packaging is often not accessible, and existing solutions like text-reading apps are insufficient for medicine-specific tasks. This project aims to create a solution that leverages pre-trained AI models to help users easily recognize medicine and understand its uses through real-time voice explanations. By employing generative AI models, this project seeks to bridge the gap between image recognition and natural language-based accessibility tools, ultimately promoting health safety and independence for disabled individuals.

2. Problem definition & challenges

The problem involves developing an AI system that accurately recognizes various medicines from images and provides real-time voice explanations. The system must reliably identify medicines from images of packaging or pills and describe essential details such as the name, purpose, dosage, and potential side effects in an accessible, voice-based format.

. The main challenges revolve around several key areas. First, medicine identification accuracy is a significant problem due to the wide variation in packaging shapes, colors, and text layouts, which are further complicated by external factors such as inconsistent lighting and partial visibility. Second, generating natural and useful descriptions presents its difficulties, as the system must simplify complex medical information while ensuring precision and clarity. Additionally, providing real-time voice explanations introduces the challenge of converting text to speech naturally and clearly, while minimizing latency. Finally, accessibility is a major concern, as the system must cater to individuals with various disabilities, requiring a user-friendly interface and strict adherence to accessibility standards.

.

3. Related Works

Research utilizing CLIP (Contrastive Language-Image Pre-training) has demonstrated its effectiveness in linking images with textual descriptions, making it highly applicable for tasks such as object recognition and medical image classification. These studies highlight CLIP's potential to recognize medicine images and assign appropriate labels accurately. Similarly, research on LLaMA (Large Language Model Meta AI) has focused on its capabilities in natural language generation, making it an ideal tool for generating clear and accurate explanations of recognized medicines. In addition, advancements in text-to-speech (TTS) technologies, like Tacotron and WaveNet, have been extensively studied for converting text into natural-sounding speech, ensuring accessibility through voice outputs. These models, proven across various domains, provide a solid foundation for the proposed project. .

4. Datasets

Focusing to refer to the public benchmark datasets such as the Pill Image Recognition Challenge Dataset will be utilized, as it contains a diverse set of images of different pill types and packaging, making it well-suited for the medicine recognition task. Additionally, the Mediqa dataset, which provides medical question-answer pairs, may be used to support the natural language generation component by helping to train the system to generate accurate medicine descriptions. If these datasets prove insufficient, a custom dataset will be collected by photographing local medicine packaging under various lighting conditions and angles, addressing real-world variability and ensuring recognition of regional brands or generic medications.

5. State-of-the-art methods and baselines

CLIP stands as a state-of-the-art model for image-text alignment tasks, consistently outperforming traditional image classifiers across various benchmarks by effectively associating images with natural language descriptions. For baseline comparisons, traditional CNNs such as ResNet and EfficientNet will be used, as they typically achieve high accuracy in image recognition tasks but require substantial

labeled data and may not generalize as effectively to unseen data as CLIP. On the language generation side, LLaMA has set a new standard by generating coherent and contextually appropriate text from minimal input, surpassing earlier models like GPT-3 in producing concise explanations. As a baseline, simpler models like RNN-based systems or GPT-2 will be used for comparison, though they tend to generate less coherent and relevant descriptions than advanced models like LLaMA.

6. Schedule

- **Weeks 1-2**: Learn and find related datasets
- **Weeks 3-4**: Start designing user interface model
- **Weeks 5-6**: Create and fine-tuning the model
- **Weeks 7-8**: Add features to model and do evaluation
- **Weeks 9-10**: Test the system and complete the report

7. References

- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Joulin, A. (2023). LLaMA: *Open and efficient foundation language models*.
- 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*.