ASSISTING PRIMARY SCHOOL STUDENT'S LEARNING THROUGH IMAGE: AN VIETNAMESE VISUAL QUESTION ANSWERING VIA SPEECH SYSTEM

Professor: PhD. Duy Le Dinh
Quan Hoang Ngoc – 22521178
University of Information Technology, VNU-HCM

Online education

- In recent years, formal education in Vietnam has undergone a significant transformation, is powered by the rapid advancement of technology, and is intensely accelerated by the COVID-19 pandemic, which led to a revolution in online education.
- However, primary school students encounter some challenges, and diverse difficulties in this novel environment.

Visual content

- Visual content is appealing, attention-grabbing, and enhances retention, aligning well with the learning preferences of many students.
- This approach is particularly beneficial for elementary school children, as it can foster curiosity, creativity, and improve their ability to understand the world around them.

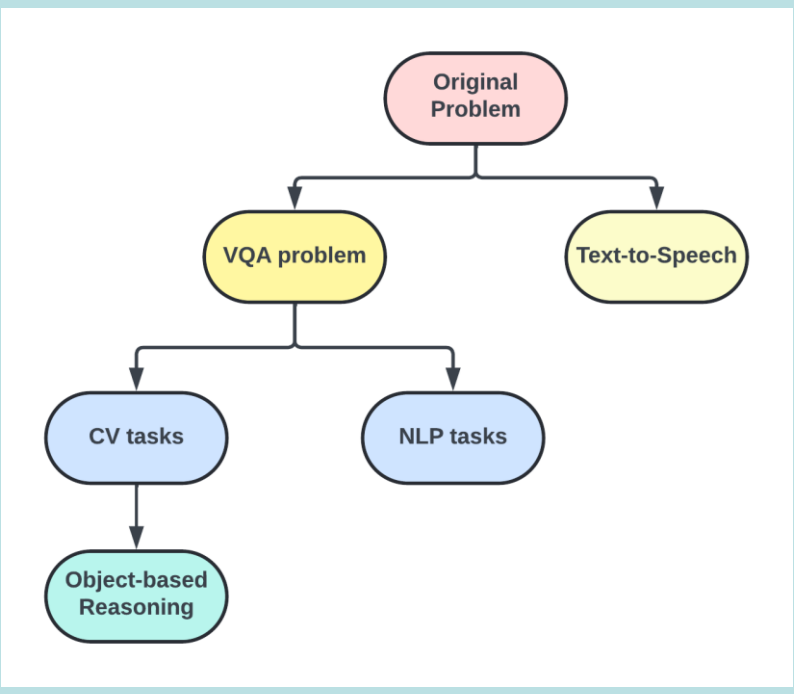
Speech-VQA system

- An education-based system for Vietnamese primary school children's learning through image-based visual content.
- Broaden their knowledge independently through visual content in a user-friendly manner, minimizing the need for excessive reading and typing by giving answers in spoken Vietnamese.
- Enhance engagement, interaction, and power for children's learning ability by not only enabling them to understand the picture and objects but also develop knowledge based on the question asked, thereby fostering curiosity, and creativity about the world around for children.

The task overview

- **Dataset Creation**: Building the **EduViVQA dataset** by collecting Vietnamese language questions paired with images from primary school textbooks and supplementary learning materials.
- **Model Development**: Designing and training **a novel VQA architecture** using the EduViVQA dataset, capable of answering diverse questions related to primary school curriculum images in Vietnamese.
- **Platform Deployment**: Creating an online web platform to implement the system. The platform will provide answers in spoken Vietnamese using Text-to-Speech models, enhancing user engagement and interaction with children.

Decomposition Tree



VQA problem

- The VQA system leverages both Computer Vision (CV) for image understanding and Natural Language Processing (NLP) for question comprehension.
- It utilizes an **Object-based Reasoning** approach, focusing on extracting object features and relationships within images to answer questions based on object properties. This approach aims to improve reasoning capabilities by recognizing objects and their interactions, leading to more accurate answers.

Text-to-Speech

- **Text-to-Speech**, which is a process of changing text into speech, can be viewed as an end-to-end seq2seq problem where textual answers are converted from a sequence of words to a sequence of audio samples.
- We aim to enhance engagement and interaction with children by providing answers in spoken Vietnamese, thereby increasing the appeal and effectiveness of the system.

EduViVQA dataset

- **EduViVQA dataset** to tailor the system to the Vietnamese educational context. This dataset will consist of Vietnamese language questions paired with images sourced from primary school textbooks and supplementary learning materials such as "Kết nối tri thức" and "Cánh diều".
- **Data analysis** identifies common image types and questions, to gain deeper insights into the dataset. These insights are utilized to create a preliminary question bank, which will suggest questions for children to select from, to enhance children's visual learning.