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

UIT – VNU HCMC
Scientific Research Methodology
Professor: PhD. Duy Le Dinh
Quan Hoang Ngoc – 22521178

Information

- Class: CS519.O21.KHTN
- Link Github: https://github.com/QuanHoangNgoc/CS519.O21.KHTN
- Link YouTube video: https://youtu.be/wCOo6yuM6F8
- Full name: Quan Hoang Ngoc

Introduction

- 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**.

Primary school students: encounter some challenges, and diverse difficulties in this novel environment.

Learning through visual content: appealing, attention-grabbing, foster curiosity, and creativity of children.

Introduction

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.
- 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.
- "How can we develop a Vietnamese Speech VQA system to enhance the learning capabilities and knowledge of Vietnamese primary school students through image-based visual content?"

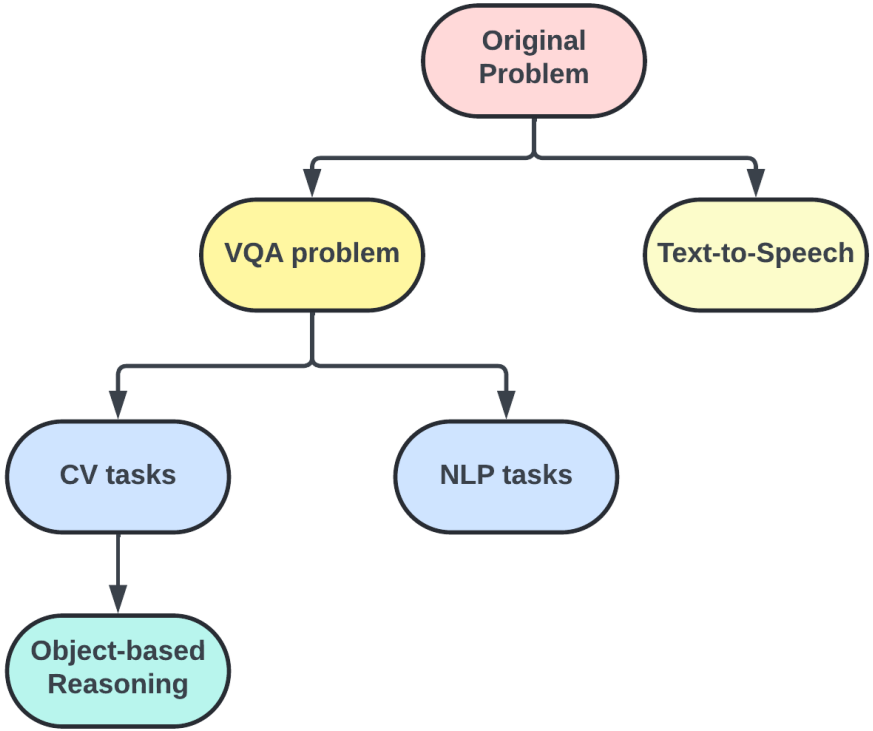
The tasks

- In the first step, we aim to build the **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".

- Following this, our crucial aim is delve into the exploration and development of **a novel VQA architecture** based on the EduViVQA dataset to *efficiently and accurately respond* to a diverse array of questions about images within the primary school curriculum, all in the Vietnamese language. [evaluated by accuracy and response time metrics]

- Lastly, create an online web platform to implement our system. Through this deployment, we aim to enhance engagement and interaction with children by providing answers in spoken Vietnamese, thereby increasing the appeal and effectiveness of the system. To achieve this, we learn about Text-to-Speech models. [evaluated by a survey testing]

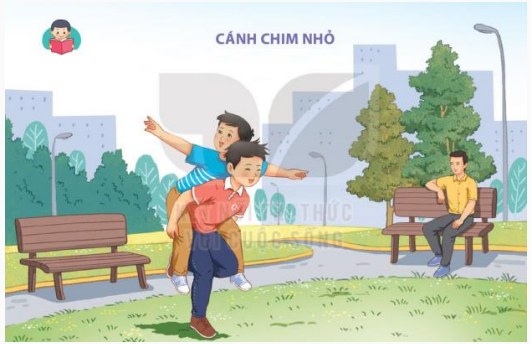
Methodology



Methodology

EduViVQA dataset

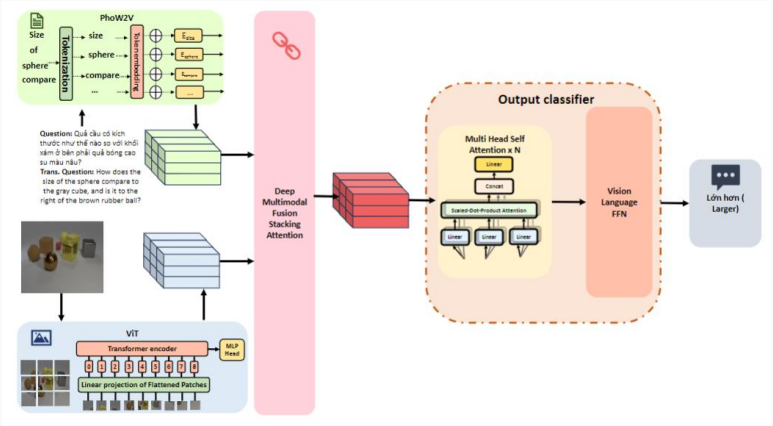
- Image source from primary school textbooks and supplementary learning materials: *"Kết nối tri thức" and "Cánh diều"*. Annotate each image and formulate questions pertaining to the content within each image.
- Data analysis identifies common image types and questions, to gain deeper insights into the dataset. These insights are utilized to create a preliminary question bank, which will suggest questions for children to select from, leading to enhance children's learning experience.



Methodology

VQA approach

- One of the keys to improving the correct answering ability of VQA models is the architecture and training strategy.
- It utilizes an **Object-based Reasoning** approach, focusing on extracting object features and relationships within images to answer questions based on object properties.
- We hypothesize that recognizing objects in the image and the relationships between them can help the model improve its ability to reason about the content of the image, thereby providing more accurate answers.



Expectation

The methodology is designed to achieve positive outcomes on various fronts:

- The **EduViVQA dataset**, with over 2,000 images and 8,000 question-answer pairs (QAs), enables tailored adaptation for the Vietnamese educational context. This diverse questions aligns with primary school curriculum topics for effective system evaluation.

- System evaluation will focus on accuracy and response time metrics. Accuracy, reflecting model reliability, will be based on correct prediction (answer) ratios, while response time will gauge question answering speed. The **Object-based Reasoning** approach aims for an 83% accuracy rate with responses under 30 seconds per question.

- A survey among elementary school students will gather feedback on system effectiveness. The goal is for 80% of participants to recognize the system's value in skill and learning experience enhancement, informing future system improving.

References

- [1]. Khiem Vinh Tran, Hao Phu Phan, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen: ViCLEVR: A Visual Reasoning Dataset and Hybrid Multimodal Fusion Model for Visual Question Answering in Vietnamese. CoRR abs/2310.18046 (2023)

- [2]. Kushal Kafle, Christopher Kanan: An Analysis of Visual Question Answering Algorithms. ICCV 2017: 1983-1991

- [3]. Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Sheng Zhao, Tao Qin, Frank K. Soong, Tie-Yan Liu: Natural-Speech: End-to-End Text-to-Speech Synthesis With Human-Level Quality. IEEE Trans. Pattern Anal. Mach. Intell. 46(6): 4234-4245 (2024)