THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/wCOo6yuM6F8*
- Link slides (dạng .pdf đặt trên Github): *https://github.com/QuanHoangNgoc/CS519.O21.KHTN*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

● Họ và Tên: Hoàng Ngọc Quân ● MSSV: 22521178 	● Lớp: CS519.O21.KHTN ● Tự đánh giá (điểm tổng kết môn): 7.5/10 ● Số buổi vắng: 0 ● Số câu hỏi QT cá nhân: I answer fully the homework questions ● Link Github: myGitHub

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

HỖ TRỢ HỌC SINH TIỂU HỌC HỌC TẬP QUA HÌNH ẢNH:
MỘT HỆ THỐNG TRẢ LỜI CÂU HỎI TRỰC QUAN BẰNG TIẾNG VIỆT
THÔNG QUA GIỌNG NÓI

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

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

TÓM TẮT *(Tối đa 400 từ)*

In the last 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 learning. Although this approach brings various advantages, several groups of learners still encounter challenges in learning effectively on screens. Primary school students, in particular, are a susceptible group, when adapting to learning in this novel environment. To address these issues, this research study focuses on developing a Vietnamese education-based Speech-VQA system to enhance the learning capabilities and knowledge of Vietnamese primary school students through image-based visual content. The proposed system empowers children between the ages of five and eleven to broaden their knowledge independently in a user-friendly manner, and engagement platform. The integration of Visual Question Answering (VQA) and Text-to-Speech in the developed system offers a promising avenue for enhancing learning experiences, where learners can import an image, choose a question from a predefined list in Vietnamese or import the question as input, and receive a spoken answer as output.

GIỚI THIỆU *(Tối đa 1 trang A4)*

In recent years, formal education in Vietnam has undergone a significant transformation, with schools and institutes of all levels rapidly adopting online learning to reduce the reliance on traditional in-person classes. Such prevalence of online learning is powered by the rapid advancement of technology and intensely accelerated by the COVID-19 pandemic. While this approach brings various advantages such as increasing accessibility to learning materials and higher convenience, several groups of learners still encounter challenges in learning effectively on screens.

Primary school students, in particular, are a susceptible group, as they may easily face distraction, reduced concentration, and lack of interest when studying on electronic devices like laptops or smartphones.

To address these issues, engaging children with content that captures their focus and interest is crucial, with visual materials such as images, pictures, and graphics proving especially effective in this regard. 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.

Therefore, it is imperative to develop an educational system that empowers children between the ages of five and eleven to broaden their knowledge independently through visual content in a user-friendly manner, minimizing the need for excessive reading and typing. The integration of Visual Question Answering (VQA) [2] and Text-to-Speech [3] models offers a promising avenue for enhancing learning experiences. By combining these technologies to create a Speech-VQA system, learners can import an image, choose questions from a predefined list in Vietnamese or import them *as input*, and receive spoken answers *as output*, thereby improving accessibility and engagement. The Speech-VQA system is powerful for improving children's learning ability by not only enabling them to understand the picture and objects but also giving them knowledge based on the question asked. Besides, it also increases engagement and interaction with children by giving answers in spoken Vietnamese. All of the observations mentioned above serve as catalysts driving our exploration of a natural question:
"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?"

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

To address the research question at hand, our tasks will unfold as follows:

- 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 to 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]

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Solving the problem for the VQA system can be known as an integration for both Computer Vision (CV) and Natural Language Processing (NLP) tasks. While Computer Vision is used for image understanding and teaching machines how to see, communication between humans and machines and teaching machines how to read are concerned with Natural Language Processing.
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. Thus, a Speech-VQA system can be decomposed into the following two independent core problems: VQA and Text-to-Speech, shown by *Figure 1*. The Text-to-Speech serves as a post-processing model to increase the interaction and usefulness of the system for children, while tailoring the VQA system to the Vietnamese educational context poses a significant research challenge that requires focused attention.

To be able to do this, in the first step, we examine and estimate the scope of this task to have the most profound understanding. Based on that, we build the EduViVQA dataset. The image sources we plan to choose are primary school textbooks and supplementary learning materials: "Kết nối tri thức" and "Cánh diều", focusing on visual-rich subjects such as natural and social sciences. After collecting image sources, we annotate each image and formulate questions pertaining to the content within each image, as well as the relationships among objects depicted. Additionally, we conduct data analysis to gain deeper insights into the dataset, identifying prevalent groups of images and common questions to enhance children's learning through visuals. These insights are utilized to create a preliminary question bank, which will suggest questions for children to select from, thereby facilitating their engagement and learning experience.

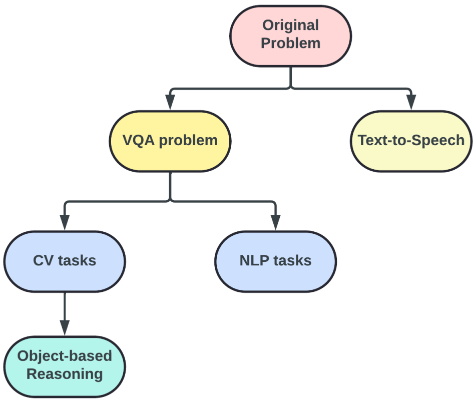

Figure 1: Problem Decomposition of Speech-VQA system

One of the keys to improving the correct answering ability of VQA models is the architecture and training strategy. VQA models will be learned, researched, and experimented with, paying special attention to the PhoViT [1], a novel method proposed by a group of authors from the University of Information Technology (UIT) - VNU HCMC. This architecture is segmented into four principal components: the Image Embedding module for assimilating visual information, the Question Embedding module for textual integration, the Multimodal Fusion module for amalgamating the extracted features, and the Classifier layer to process the fused features representation and predict the most likely answer to the question. The model will be fine-tuned on the EduViVQA dataset that we have built to tailor the VQA system to the Vietnamese educational context. Moreover, we also develop the model towards *Object-based Reasoning*, which allows the model to extract features and relationships between objects in the image to answer the question based on these objects' features. 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.

Finally, we search for an appropriate Text-to-Speech model in spoken Vietnamese to set up and integrate into our system. Then, we use client-server frameworks such as FastAPI, and Flutter to build the system as an online web platform, deploy the trained model operation as a server, and also design an attractive and friendly interface for children. 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.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

The proposed methodology aims to achieve positive outcomes in several key aspects:

1. The EduViVQA dataset, consisting of over 2,000 images and 8,000 question-answer pairs (QAs), is sufficiently large to customize the system for the Vietnamese educational context. This dataset offers a diverse range of questions aligned with the primary school curriculum for each image, enabling an objective evaluation of the system's effectiveness.

2. The system's performance will be assessed using two primary metrics: accuracy and response time. Accuracy, reflecting the model's reliability, is calculated based on the ratio of correct predictions to total predictions (answers). Meanwhile, response time will measure the system's speed in answering each question. It is anticipated that employing the Object-based Reasoning approach will result in an average accuracy rate of approximately 83% and a response time under 30 seconds per question.

3. Additionally, a survey will be conducted among a randomly selected group of elementary school students to gather feedback. The aim is to ascertain student perceptions of the system's effectiveness, with the hoped-for outcome that 80% of the surveyed students will acknowledge the system's value in enhancing their skill set and learning experience. These insights will further inform the system's refinement and potential future enhancements.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[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:
NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality.
IEEE Trans. Pattern Anal. Mach. Intell. 46(6): 4234-4245 (2024)