

BỘ DỮ LIỆU ĐA THỂ THỨC TIẾNG VIỆT VỀ HIỂU BIẾT VÀ LUẬN LÝ KHOA HỌC TỰ NHIÊN (VMNSU)

Đặng Hữu Phát

Trường Đại học Công nghệ thông tin TP. Hồ Chí Minh

Phan Hoàng Phước

Trường Đại học Công nghệ thông tin TP. Hồ Chí Minh

What ?

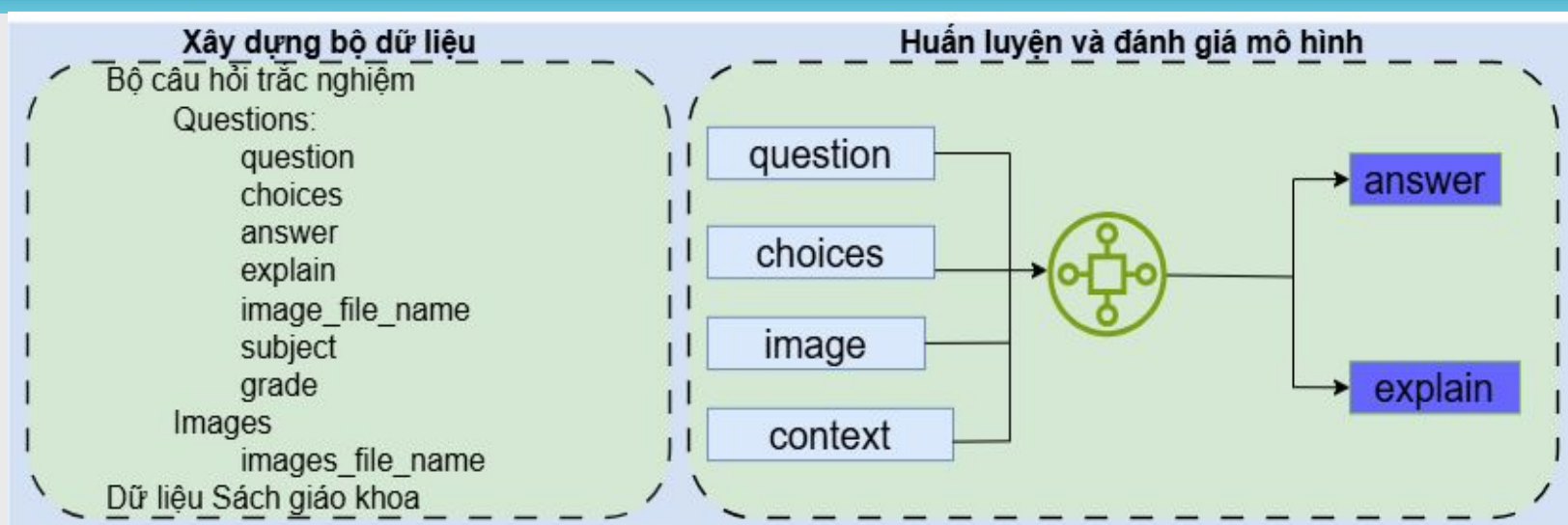
Giới thiệu bộ dữ liệu đa thể thức tiếng việt nhằm đánh giá khả năng hiểu biết và lý luận của các mô hình, cụ thể hơn:

- Bộ dữ liệu VMNSU đa thể thức tiếng việt trong lĩnh vực khoa học tự nhiên gồm câu hỏi, hình ảnh, lựa chọn trắc nghiệm và nội dung sách giáo khoa cấp THPT.
- Sử dụng các mô hình hiện nay để đánh giá dựa trên tập dữ liệu.
- Đề xuất phương pháp mới để giải quyết bài toán.

Why ?

- Bài toán VQA (Visual Question Answering) là một trong những thách thức mới, nhận được rất nhiều sự quan tâm ở lĩnh vực multimodal trong những năm gần đây.
- Tuy nhiên, lĩnh vực này vẫn còn rất mới mẻ và thiếu hụt các bộ dữ liệu chất lượng cao, đặc biệt là ở ngôn ngữ tiếng việt.
- Chúng tôi xây dựng bộ dữ liệu đa thể thức tiếng Việt về hiểu biết và luận lý khoa học tự nhiên (VMNSU) để đóng góp cộng đồng nghiên cứu ở Việt Nam.

Overview



Description

1. Xây dựng bộ datasets

Chia ra thành 3 giai đoạn xử lý chính:

- Giai đoạn 1:** Tìm kiếm và khảo sát nguồn tài nguyên có sẵn như các trang học tập trên internet.
- Giai đoạn 2:** Cài đặt chương trình thu thập dữ liệu thô từ các trang web. Dùng công cụ OCR có sẵn để trích xuất nội dung file PDF sách giáo khoa và thủ công tách các phần nội dung.
- Giai đoạn 3:** Tiền xử lý dữ liệu thành format thống nhất và loại bỏ những dữ liệu gây nhiễu, bị sai và trùng lặp trong quá trình thu thập. Chọn lọc những dữ liệu dựa trên tính đầy đủ, nhất quán và đại diện


BIOLOGY	
Question: Trong mô hình điều hòa Operon Lac được mô tả như hình bên dưới. Hai gen nào sau đây có số lần phiên mã khác nhau?	
Subject: Sinh học	Grade: Thi thử THPT Quốc gia
Choices:	
A. Gen Z và gen điều hòa.	B. Gen Z và gen A
C. Gen Z và gen Y.	D. Gen Y và gen A.
Explain: Chọn đáp án A. Gen điều hòa và các gen cấu trúc Z, Y, A có số lần phiên mã khác nhau. Các gen Z, Y, A có số lần phiên mã bằng nhau.	

Figure 2. Phần tử mẫu trong tập dữ liệu

2. Đánh giá trên các mô hình hiện nay trên bộ dữ liệu

- Khảo sát và lựa chọn các mô hình mã nguồn mở tiên tiến như BLIP-2, InstructBLIP, OpenFlamingo2-9B, LLaVA, FLAN-T5-XXL, ...
- Huấn luyện các mô hình trên bộ dữ liệu đã xây dựng và đánh giá hiệu suất dựa trên tiêu chí **độ chính xác, khả năng suy luận và thời gian xử lý**.

3. Đề xuất phương pháp mới

- Phân tích những hạn chế và ưu điểm của các phương pháp hiện tại, từ đó đề xuất cho hướng tiếp cận mới của bài toán.
- Thử nghiệm sử dụng BERT, XLNet cho text encoder và ViT, CLIP cho image encoder
- Thực nghiệm và đánh giá phương pháp, hướng tiếp cận mới để xác định khả năng giải quyết bài toán suy luận đa thể thức

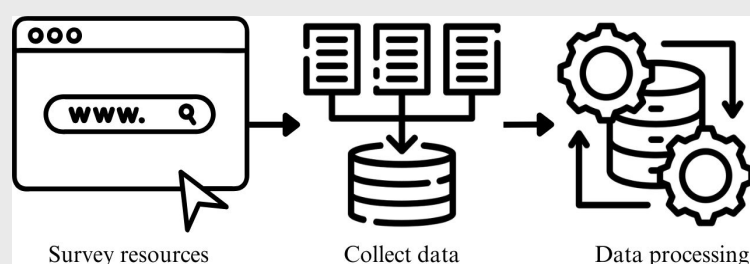


Figure 1. Quy trình thu thập và xử lý dữ liệu.

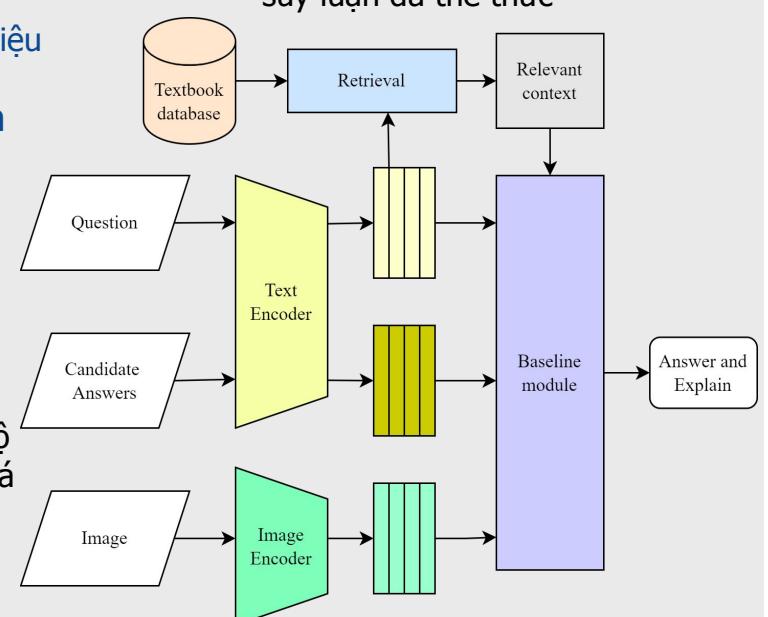


Figure 3. Mô tả tổng quan phương thức hoạt động của phương thức đề xuất.