THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://www.youtube.com/watch?v=kIEodl1Kpog*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/LeNguyenAnhKhoa/CS519.Q11/blob/main/RESEARCH_AND_DEVELOPMENT_OF_AN_ADAPTIVE_ITERATIVE_FRAMEWORK_FOR_MULTIMODAL_LONG-DOCUMENT_QUESTION_ANSWERING.pdf*
- *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: Lê Nguyễn Anh Khoa
• MSSV: 23520742
 
	• Lớp: CS519.Q11
	• Tự đánh giá (điểm tổng kết môn): 9.5/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: 8
	• Số câu hỏi QT của cả nhóm: 16
	• Link Github: https://github.com/LeNguyenAnhKhoa/CS519.Q11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đề tài
	  ○ Làm poster đề tài
	  ○ Làm video YouTube

• Họ và Tên: Cáp Kim Hải Anh • MSSV: 23520036	• Lớp: CS519.Q11 • Tự đánh giá (điểm tổng kết môn): 9.5/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 8

	• Số câu hỏi QT của cả nhóm: 16
	• Link Github: https://github.com/LeNguyenAnhKhoa/CS519.Q11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Làm slides đề tài
	  ○ Viết đề cương nghiên cứu

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NGHIÊN CỨU VÀ PHÁT TRIỂN KHUNG LÀM VIỆC LẶP THÍCH ỨNG CHO BÀI TOÁN HỎI ĐÁP TRÊN TÀI LIỆU DÀI ĐA PHƯƠNG THỨC

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

RESEARCH AND DEVELOPMENT OF AN ADAPTIVE ITERATIVE FRAMEWORK FOR MULTIMODAL LONG-DOCUMENT QUESTION ANSWERING

TÓM TẮT *(Tối đa 400 từ)*

Trong bối cảnh bùng nổ dữ liệu số, các tài liệu dài và đa phương thức (không chỉ chứa văn bản mà còn đan xen hình ảnh và bảng biểu) là rất phổ biến, kéo theo đó, nhu cầu tự động hóa việc trả lời câu hỏi (Question Answering - QA) kèm trích xuất thông tin từ các tài liệu phức tạp này đang trở nên cấp thiết. Tuy nhiên, các phương pháp hỏi đáp hiện hành, điển hình là RAG (Retrieval-Augmented Generation) đơn lượt, thường bộc lộ hạn chế khi xử lý các tài liệu dài và đa phương thức. Vấn đề cốt lõi nằm ở việc các mô hình này thường truy xuất thông tin một cách thụ động dựa vào sự tương đồng bề mặt ngữ nghĩa, dẫn đến bỏ sót các thông tin ngữ cảnh quan trọng nằm rải rác hoặc ẩn sâu trong các định dạng đa phương thức.

Xuất phát từ thực tiễn trên, đề tài thực hiện nghiên cứu và phát triển khung làm việc lặp thích ứng (Adaptive Iterative Framework) với mục tiêu khả năng suy luận đa bước trên tài liệu đa phương thức. Ý tưởng chủ đạo của nghiên cứu là từ quy trình truy xuất tĩnh chuyển sang một cơ chế động, nơi hệ thống có thể tự đánh giá độ thiếu hụt thông tin và chủ động tìm kiếm dữ liệu bổ sung. Cụ thể, chúng tôi dự kiến áp dụng lý thuyết tối ưu hóa thông tin để định lượng độ lợi thông tin (Information Gain) tại mỗi bước lập luận, kết hợp với việc sử dụng mô hình ngôn ngữ lớn (LLM) như một tác nhân điều phối thông minh. Tác nhân này đảm nhiệm phân tích câu hỏi, sinh ra các truy

vấn con (sub-queries) để lấp đầy khoảng trống kiến thức, và tinh chỉnh kết quả truy xuất thông qua cơ chế phản hồi liên tục. Kết quả nghiên cứu kỳ vọng sẽ tạo ra một giải pháp QA mạnh mẽ, có khả năng xử lý các câu hỏi phức tạp đòi hỏi sự tổng hợp thông tin từ nhiều nguồn và định dạng khác nhau, khắc phục điểm yếu của các hệ thống truy xuất đơn lượt truyền thống.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong kỷ nguyên chuyển đổi số, nhu cầu khai thác thông tin từ các tài liệu điện tử ngày càng trở nên cấp thiết. Tuy nhiên, một vấn đề lớn đang tồn tại là sự phức tạp của các tài liệu thực tế (tạp chí khoa học, hồ sơ kỹ thuật, báo cáo tài chính). Các tài liệu này thường không những là văn bản thuần túy mà còn chứa đựng lượng lớn thông tin dưới dạng biểu đồ, bảng và hình ảnh minh họa, được trình bày đan xen nhiều trang. Bài toán đặt ra là làm thế nào để xây dựng một hệ thống có khả năng đọc hiểu toàn diện và trả lời chính xác các câu hỏi dựa trên nguồn dữ liệu hỗn hợp này.

Hiện nay, các mô hình hỏi đáp QA truyền thống thường sử dụng kỹ thuật RAG đơn lượt. Phương pháp này hoạt động dựa trên nguyên lý tìm kiếm các đoạn văn bản có từ khóa tương đồng với câu hỏi. Tuy nhiên, thông qua quá trình phân tích và tư duy phản biện, chúng tôi nhận thấy hướng tiếp cận này bộc lộ hạn chế lớn khi đối mặt với các câu hỏi đòi hỏi suy luận đa bước. Ví dụ, khi thông tin trả lời nằm rải rác: một phần dữ liệu nằm ở biểu đồ trang 6, phần còn lại nằm ở chú thích trang 19. Các hệ thống hiện tại thường thất bại trong việc kết nối các manh mối rời rạc này, dẫn đến câu trả lời thiếu chính xác hoặc bị ảo giác (hallucination).

Xuất phát từ thực tiễn và những hạn chế kỹ thuật nêu trên, chúng tôi lựa chọn đề tài này với mục tiêu nghiên cứu và phát triển một khung làm việc lặp thích ứng (Adaptive Iterative Framework). Thay vì truy xuất thụ động, hệ thống đề xuất sẽ áp dụng tư duy máy tính (Computational Thinking) để mô phỏng quy trình suy luận của con người: tự đặt câu hỏi phụ, tìm kiếm bằng chứng, đánh giá mức độ đủ của thông

tin và tiếp tục tìm kiếm nếu cần thiết.

Để làm rõ phạm vi và tính khả thi của đề tài, chúng tôi xác định các thành phần đầu vào và đầu ra của hệ thống như sau:

- Đầu vào (Input): Hệ thống tiếp nhận hai thành phần chính:
    - Tài liệu nguồn: Một hoặc nhiều tập tin định dạng PDF hoặc hình ảnh (slide, scan) có độ dài lớn (ví dụ: dài 20-50 trang), chứa nội dung hỗn hợp gồm văn bản, biểu đồ và bảng.
    - Câu truy vấn: Một câu hỏi tự nhiên từ người dùng, có thể yêu cầu thông tin chi tiết (ví dụ: "Tỷ lệ tăng trưởng doanh thu trong biểu đồ trang 15 so với số liệu dự báo ở trang 3 là bao nhiêu?").
- Đầu ra (Output):
    - Câu trả lời văn bản: Kết quả chính xác, súc tích được tổng hợp từ quá trình suy luận (ví dụ: "Tỷ lệ tăng trưởng là 15%, dựa trên doanh thu thực tế 50 tỷ ở trang 15 và dự báo 43.5 tỷ ở trang 3").
    - Minh chứng trích dẫn: Hệ thống chỉ rõ nguồn gốc thông tin được lấy từ trang nào, hình ảnh/đoạn văn nào để người dùng có thể kiểm chứng.

Nghiên cứu này sẽ kỳ vọng không chỉ mang ý nghĩa khoa học trong việc đóng góp một giải pháp hiệu quả cho bài toán trích xuất thông tin tự động, khắc phục điểm yếu của các mô hình RAG tĩnh hiện nay mà còn có tính ứng dụng cao trong các lĩnh vực phân tích như tài chính, kinh tế và giáo dục.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

Để giải quyết các thách thức trong việc trích xuất và tổng hợp thông tin từ tài liệu đa phương thức, đề tài tập trung thực hiện 03 mục tiêu cụ thể sau:

1. **Xây dựng cơ sở lý thuyết và mô hình hóa toán học cho bài toán Hỏi đáp đa phương thức:**

- Nghiên cứu và hệ thống hóa các phương pháp hiện hành về Multimodal RAG.
- Thiết lập công thức toán học cho bài toán tối ưu hóa thông tin tương hỗ (Mutual Information Maximization) nhằm định lượng chính xác độ lợi thông tin (Information Gain) và độ không chắc chắn trong quá trình truy xuất dữ liệu từ văn bản và hình ảnh.

2. **Thiết kế và phát triển Khung làm việc lặp thích ứng (Adaptive Iterative Framework):**

- Xây dựng thuật toán truy xuất động có khả năng tự động phân tích câu hỏi phức tạp thành các truy vấn con (sub-queries).
- Phát triển module tích hợp mô hình ngôn ngữ lớn (LLM) đóng vai trò tác nhân đánh giá để kiểm tra tính đầy đủ của thông tin sau mỗi bước truy xuất, từ đó quyết định dừng lại hoặc tiếp tục tìm kiếm thêm bằng chứng.
- Cài đặt cơ chế tối ưu hóa hàm mất mát InfoNCE để cải thiện độ chính xác việc khớp nối giữa câu truy vấn và nội dung đa phương thức trong tài liệu.

3. **Thực nghiệm, đánh giá và tối ưu hiệu suất hệ thống:**

- Triển khai hệ thống trên các bộ dữ liệu chuẩn về tài liệu đa phương thức như SlideVQA hoặc MMLongBench-Doc.
- Đo lường và so sánh hiệu quả của mô hình đề xuất với các phương pháp RAG truyền thống (như ColPali, VisRAG) dựa trên các chỉ số định lượng: Độ chính xác (Accuracy), F1-Score và số bước lặp trung bình cần thiết để đưa ra câu trả lời đúng.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Nghiên cứu cơ sở lý thuyết và xây dựng mô hình toán học

Mục tiêu:

- Hệ thống hóa các kiến thức nền tảng về Multimodal RAG và các hạn chế của phương pháp truy xuất đơn lượt.

- Xây dựng công thức toán học định lượng sự cân bằng giữa độ lợi thông tin (Information Gain) và độ không chắc chắn trong quá trình truy xuất.

Sản phẩm dự kiến:

- Báo cáo tổng quan tài liệu về các phương pháp hỏi đáp trên tài liệu.
- Mô hình toán học hoàn chỉnh mô tả bài toán tối ưu hóa thông tin tương hỗ (Mutual Information Maximization).

Phương pháp thực hiện:

1. **Thu thập và phân tích tài liệu:** Tìm kiếm và đọc hiểu các bài báo khoa học liên quan đến RAG, InfoNCE, và các kiến trúc LLM đa phương thức (như GPT-4V, Gemini, LayoutLM) trên các cơ sở dữ liệu uy tín (Google Scholar, arXiv, ACL Anthology).
2. **Mô hình hóa bài toán (Abstraction):**

- Định nghĩa bài toán dưới dạng tối ưu hóa hàm mục tiêu, trong đó đầu vào là tập hợp các trang tài liệu hỗn hợp $D = \{p_1, p_2, ..., p_N\}$ và câu truy vấn Q.
- Thiết lập công thức tính Entropy H(.) để đo lường độ không chắc chắn của thông tin được trích xuất.
- Áp dụng lý thuyết Information Bottleneck để xây dựng hàm mất mát, giúp xác định điểm dừng tối ưu khi lượng thông tin thu được đã đủ để trả lời câu hỏi.

Nội dung 2: Thiết kế và phát triển khung làm việc lặp thích ứng

Mục tiêu:

- Hiện thực hóa thuật toán truy xuất động, cho phép hệ thống tự động lập kế hoạch tìm kiếm thông tin qua nhiều bước.
- Xây dựng các module chức năng: Module phân hóa thông tin, Module truy xuất hướng dẫn và Module phản hồi.

Sản phẩm dự kiến:

- Mã nguồn hoàn chỉnh của hệ thống.
- Bộ Prompt được tối ưu hóa cho tác nhân LLM.

Phương pháp thực hiện:

1. **Thiết kế kiến trúc hệ thống:**

- Sử dụng mô hình Client-Server, trong đó Server xử lý logic truy xuất và client là giao diện tương tác người dùng.
- Lựa chọn công nghệ: Python (ngôn ngữ chính), PyTorch (xây dựng mô hình học sâu), LangChain hoặc LlamaIndex (quản lý luồng tác nhân LLM).

2. **Xây dựng module phân hóa thông tin (Information Differentiation):**

- Lập trình quy trình suy nghĩ cho LLM: Trước khi tìm kiếm, LLM phải phân tích câu hỏi Q và xác định những thông tin còn thiếu.
- Cài đặt cơ chế lưu trữ trạng thái để ghi nhớ những gì đã tìm thấy ở các bước trước đó, tránh việc tìm kiếm lặp lại vô ích.

3. **Cài đặt thuật toán Truy xuất hướng dẫn bởi InfoNCE:**

- Tích hợp các mô hình Embedding đa phương thức (như CLIP hoặc ColBERT) để chuyển đổi văn bản và hình ảnh thành vector đặc trưng.
- Lập trình hàm mất mát InfoNCE để xếp hạng các trang tài liệu dựa trên mức độ đóng góp thông tin cho câu trả lời, thay vì chỉ dựa trên độ tương đồng từ khóa.

4. **Phát triển tác nhân phản hồi:**

- Thiết kế các mẫu câu lệnh (Prompt template) để hướng dẫn LLM đóng vai trò người giám khảo. Ví dụ: *"Dựa trên dữ liệu trang 5 vừa tìm được, đã đủ để trả lời câu hỏi chưa? Nếu chưa, cần tìm thêm thông tin gì?"*.
- Xây dựng cơ chế sinh truy vấn phụ (sub-query) để hệ thống tự động thực hiện

các lệnh tìm kiếm tiếp theo.

Nội dung 3: Thực nghiệm, đánh giá và tối ưu hóa

Mục tiêu:

- Kiểm chứng độ chính xác và tính ổn định của hệ thống trên dữ liệu thực tế.
- So sánh hiệu năng với các phương pháp tiên tiến hiện nay để chứng minh tính ưu việt.

Sản phẩm dự kiến:

- Bảng số liệu kết quả thực nghiệm (Accuracy, F1-Score).
- Phân tích các trường hợp cụ thể minh họa khả năng suy luận của mô hình.

Phương pháp thực hiện:

1. **Chuẩn bị dữ liệu:**

- Sử dụng các bộ dữ liệu chuẩn như SlideVQA (bộ câu hỏi trên slide thuyết trình) và MMLongBench-Doc (tài liệu dài đa phương thức)
- Tiền xử lý dữ liệu: Trích xuất văn bản (OCR), cắt vùng ảnh, chuẩn hóa định dạng đầu vào.

2. **Thiết lập kịch bản kiểm thử:**

- Kịch bản 1: So sánh với phương pháp RAG truyền thống sử dụng truy xuất đơn lượt.
- Kịch bản 2: Đánh giá khả năng xử lý câu hỏi suy luận đa bước (Multi-hop Reasoning) – câu hỏi yêu cầu liên kết thông tin từ ít nhất 2 trang khác nhau.

3. **Đánh giá và Tinh chỉnh:**

- Sử dụng các chỉ số định lượng: Độ chính xác chính xác (Exact Match), độ đo ngữ nghĩa (BERTScore, GPT-4 Judge).

- Phân tích lỗi: Xem xét các trường hợp mô hình trả lời sai để điều chỉnh lại thuật toán InfoNCE hoặc cải thiện prompt.

KẾT QUẢ MONG ĐỢI

1. Về hiệu suất kỹ thuật: Hệ thống đạt độ chính xác kỳ vọng từ 40-55% trên tập dữ liệu chuẩn, cải thiện đáng kể khả năng xử lý các câu hỏi suy luận đa bước phức tạp so với các mô hình RAG đơn lượt hiện hành.
2. Phân tích lỗi: Đánh giá các trường hợp thất bại để hiểu rõ giới hạn của phương pháp xấp xỉ thông tin.
3. Đề xuất được quy trình chuẩn hóa áp dụng lý thuyết tối ưu hóa thông tin vào bài toán truy xuất tài liệu.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao, "ReAct: Synergizing reasoning and acting in language models," *arXiv preprint arXiv:2210.03629*, 2022.

[2] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo, "ColPali: Efficient document retrieval with vision language models," *arXiv preprint arXiv:2407.01449*, 2024.

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[4] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito, "SlideVQA: A dataset for document visual question answering on multiple images," in *AAAI*, 2023.

[5] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al., "VisRAG: Vision-based retrieval-

augmented generation on multi-modality documents," *arXiv preprint arXiv:2410.10594*, 2024.

[6] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al., "MMLongBench-Doc: Benchmarking long-context document understanding with visualizations," *arXiv preprint arXiv:2407.01523*, 2024.

[7] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459-9474.

[9] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang, "MM-React: Prompting ChatGPT for Multimodal Reasoning and Action," *arXiv preprint arXiv:2303.11381*, 2023.