

# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://www.youtube.com/watch?v=7nTh4VCJK6Y>
- Link slides:  
[https://github.com/Be-Tap-Code/CS519.P11/blob/main/slide\\_CS519.pdf](https://github.com/Be-Tap-Code/CS519.P11/blob/main/slide_CS519.pdf)

- Họ và Tên: Trần Minh Quân
- MSSV: 22521191



- Lớp: **CS519.P11**
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 1
- Số câu hỏi QT cá nhân: 10
- Số câu hỏi QT của cả nhóm: 10
- Link Github:  
<https://github.com/mynameuit/CS519.P11/>
- Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
  - Lên ý tưởng và thiết kế đề tài nghiên cứu
  - Viết báo cáo nội dung
  - Làm video YouTube
  - Làm slide thuyết trình
  - Làm poster báo cáo

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI

CẢI TIẾN VÀ TỐI ƯU HÓA MÔ HÌNH ĐA PHƯƠNG THỨC CHO TẠO SINH BÁO CÁO HÌNH ẢNH X-QUANG

## TÊN ĐỀ TÀI TIẾNG ANH



IMPROVING AND OPTIMIZING OF MULTIMODAL MODELS FOR AUTOMATED X-RAY REPORT GENERATION

## TÓM TẮT

Trong bối cảnh y học hiện đại, công nghệ hình ảnh y khoa, đặc biệt là hình ảnh X-quang, đã trở thành công cụ thiết yếu trong việc chẩn đoán và theo dõi tình trạng sức khỏe của bệnh nhân. Hình ảnh X-quang cung cấp thông tin quan trọng về cấu trúc và chức năng của cơ thể, tuy nhiên quy trình tạo lập và viết báo cáo cho các hình ảnh này vẫn gặp nhiều thách thức. Những thách thức này bao gồm tính phức tạp của dữ liệu hình ảnh và yêu cầu về kinh nghiệm chuyên môn cao để đảm bảo độ chính xác trong chẩn đoán. Mặc dù công nghệ học sâu đã mang lại những cải tiến đáng kể, nhưng độ tin cậy của các phương pháp tự động hóa hiện tại trong việc tạo ra báo cáo X-quang vẫn còn hạn chế.

Đề tài nghiên cứu này tập trung vào phát triển một mô hình đa phương thức mới nhằm cải thiện độ chính xác và tính mạch lạc trong việc tạo ra báo cáo X-quang tự động. Mục tiêu chính của nghiên cứu là tối ưu hóa các mô hình hiện có bằng cách áp dụng các kiến trúc tiên tiến như *Vision Transformer (ViT)* và *BioBERT*. Các kiến trúc này được kỳ vọng sẽ nâng cao khả năng học hỏi từ các bộ dữ liệu đa dạng, từ đó cải thiện độ chính xác và khả năng tổng quát của mô hình.

Nghiên cứu cũng sẽ tập trung vào việc khám phá các phương pháp đánh giá chất lượng cho báo cáo y khoa tự động. Việc xây dựng một bộ công cụ đánh giá toàn diện là rất quan trọng, giúp xác định độ chính xác ngữ nghĩa và tính nhất quán của các báo cáo được sinh ra. Kết quả dự kiến của nghiên cứu bao gồm một hệ thống tạo báo cáo X-quang tự động với độ chính xác cao, có khả năng hỗ trợ các bác sĩ lâm sàng trong việc phân tích hình ảnh y khoa và cải thiện chất lượng chăm sóc sức khỏe cho bệnh nhân. Hệ thống này không chỉ giảm bớt gánh nặng công việc cho các bác sĩ mà còn nâng cao độ tin cậy trong chẩn đoán và điều trị bệnh.

Input Image	Recurrent Attention	Ground Truth
	<p><b>Findings:</b> The heart size and mediastinal contours appear within normal limits. No focal airspace consolidation, pleural effusion or pneumothorax. No acute bony abnormalities.</p> <p><b>Impression:</b> No acute cardiopulmonary finding.</p>	<p><b>Findings:</b> The heart size and mediastinal silhouette are within normal limits for contour. The lungs are clear. No pneumothorax or pleural effusions. The XXXX are intact.</p> <p><b>Impression:</b> No acute cardiopulmonary abnormalities.</p>
	<p><b>Findings:</b> The heart size and mediastinal silhouette are within normal limits for contour. The lungs are clear. No focal airspace consolidation. No pleural effusion or pneumothorax. Normal cardiomeastinal silhouette. Heart size is normal.</p> <p><b>Impression:</b> Clear lungs. No acute cardiopulmonary abnormality.</p>	<p><b>Findings:</b> Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. No bony abnormality. Vague density in right mid lung, XXXX related to scapular tip and superimposed ribs. Not visualized on lateral exam.</p> <p><b>Impression:</b> Vague density in right XXXX, XXXX related to scapular tip and superimposed ribs. Consider oblique images to exclude true nodule. 2. No acute cardiopulmonary abnormality.</p>

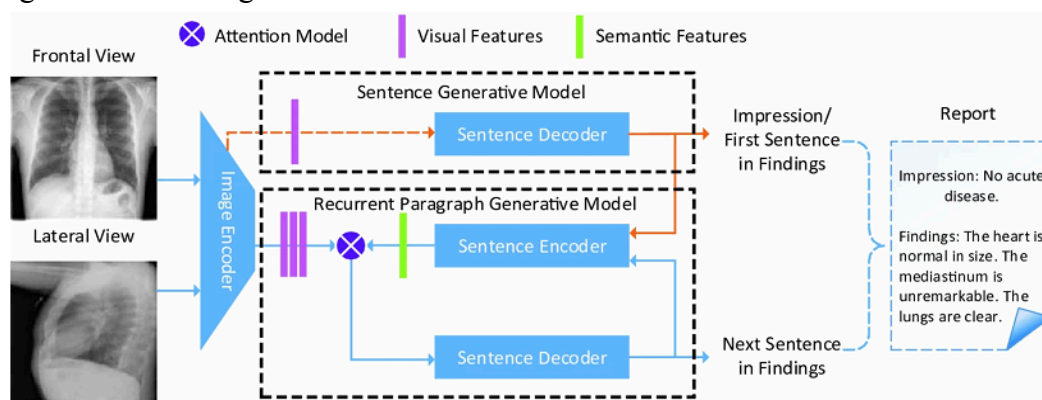
Hình 1. Ví dụ về báo cáo gốc so với báo cáo được tạo bởi recurrent attention model

## GIỚI THIỆU

Hầu hết các tài liệu hiện có liên quan đến vấn đề tạo báo cáo kết luận cho ảnh X-quang đều dựa trên công nghệ học sâu, tuân theo kiến trúc encode-decode ban đầu được sử dụng cho dịch máy. Để tạo mô tả đoạn văn, là một chuỗi rất dài, một số công trình tiên phong đã được thực hiện trong lĩnh vực chú thích hình ảnh tự nhiên, với các *mạng hồi quy phân cấp* [1] và *mạng Long Short-Term Memory (LSTM)* [2] gặp khó khăn trong việc duy trì tính mạch lạc theo ngữ cảnh trong các chuỗi dài, thường dẫn đến các vấn đề như lặp lại hoặc mất luồng ngữ nghĩa. Các khuôn khổ tạo phân cấp đã được đề xuất để giải quyết các vấn đề này bằng cách sử dụng multi-level RNN để tạo các biểu diễn paragraph-level và sentence-level. Tuy nhiên, các phương pháp này thường không khai thác hết bản chất đa phương thức của dữ liệu hình ảnh y tế hoặc gặp phải những hạn chế do thiếu các mô hình chú thích dày đặc được đào tạo trước cho hình ảnh y tế.

Bài báo “*Multimodal Recurrent Model with Attention for Automated Radiology Report Generation*” [3], đã đề xuất một mô hình kết hợp đa modal (hình ảnh và ngữ nghĩa) với cơ chế hồi tiếp và attention để tạo báo cáo chi tiết (Hình 2), bao gồm các phát hiện (findings) và kết luận chính (impression). Dựa trên kiến trúc của mô hình này, đề tài sẽ đề xuất một framework mới để tạo báo cáo X-quang tự động tập trung vào việc cải thiện tính mạch lạc, độ chính xác và khả năng thích ứng của các báo cáo được tạo. Bao gồm các giai đoạn sau:

- **Giai đoạn 1: Xử lý hình ảnh đầu vào:** Sử dụng hai góc nhìn X-quang (frontal và lateral) làm đầu vào để đảm bảo thông tin đầy đủ hơn về tình trạng của bệnh nhân. Hình ảnh đầu vào được xử lý thông qua một encoder học sâu. Đầu ra là các visual features (đặc trưng hình ảnh) đại diện cho các vùng trong hình.
- **Giai đoạn 2: Kết hợp đặc trưng ngữ nghĩa:** Các đặc trưng hình ảnh được kết hợp với các khái niệm ngữ nghĩa y khoa. Cơ chế attention [4] được áp dụng để học mối liên hệ giữa các vùng hình ảnh và các khái niệm ngữ nghĩa, giúp mô hình tập trung vào những vùng quan trọng trong ảnh.
- **Giai đoạn 3: Tạo câu và đoạn văn (Report Generation):** Bao gồm hai thành phần chính là Sentence Generative Model và Recurrent Paragraph Generative Model. Mô hình Sentence Generative Model chịu trách nhiệm tạo ra câu đầu tiên trong báo cáo, thường là phần Impression hoặc câu mở đầu của Findings. Mỗi câu trong báo cáo được tạo dựa trên câu trước đó thông qua một sentence encoder và sentence decoder. Điều này đảm bảo tính mạch lạc giữa các câu trong báo cáo.



Hình 2. Kiến trúc tổng quan của mô hình baseline

Nghiên cứu sắp tới sẽ kế thừa và cải tiến phương pháp này, tập trung vào việc nâng cao độ chính xác của báo cáo thông qua việc tích hợp các mô hình hiện đại hơn và tối ưu hóa quá trình học tập đa

phương thức (multimodal learning)

Đóng góp chính và tính mới của nghiên cứu này bao gồm:

- **Phát triển mô hình đa phương thức:** Nghiên cứu sẽ giới thiệu một mô hình đa phương thức mới kết hợp các công nghệ hiện đại như *Vision Transformer* và *BioBERT*, nhằm cải thiện khả năng hiểu và phân tích báo cáo hình ảnh X-quang.
- **Tăng cường độ chính xác:** Nghiên cứu sẽ tối ưu hóa quy trình tạo sinh báo cáo X-quang, nâng cao độ chính xác và tính mạch lạc của thông tin trong báo cáo.
- **Hệ thống tự động hóa:** Đề tài sẽ phát triển một hệ thống tự động hóa toàn diện cho việc tạo báo cáo X-quang, giúp giảm thiểu thời gian và công sức của bác sĩ trong quá trình chẩn đoán.

## MỤC TIÊU

Đề tài tập trung vào các mục tiêu sau:

- **Tăng khả năng học của mô hình từ các bộ dữ liệu đa dạng:** Bằng cách kết hợp nhiều bộ dữ liệu y khoa khác nhau để mở rộng khả năng tổng quát của mô hình, giảm thiểu tình trạng overfitting.
- **Cải thiện độ chính xác của mô hình tạo sinh báo cáo tự động:** Tối ưu hóa kiến trúc của mô hình hiện tại nhằm nâng cao độ chính xác trong việc tạo ra báo cáo tự động từ hình ảnh y khoa, giúp bác sĩ có được thông tin chính xác và nhanh chóng.
- **Khám phá các kiến trúc mô hình mới:** Áp dụng các mô hình hiện đại như *Vision Transformers (ViT)* cho phần xử lý hình ảnh và *BERT/BioBERT/...* cho phần xử lý văn bản, nhằm tối ưu hóa hiệu quả tạo báo cáo đầu ra.

## NỘI DUNG VÀ PHƯƠNG PHÁP

### Nội dung 1: Nghiên cứu và phân tích mô hình gốc

Mô hình gốc được xây dựng trên cơ sở **Multimodal Recurrent Model with Attention [3]**, kết hợp giữa các lớp RNN/LSTM với cơ chế Attention để xử lý dữ liệu hình ảnh và văn bản. Tuy nhiên, các mô hình RNN gặp phải hạn chế trong việc xử lý mối quan hệ dài hạn và khả năng tính toán song song, dẫn đến giảm hiệu suất và tốc độ xử lý với dữ liệu lớn.

#### Phương pháp thực hiện:

- Phân tích kiến trúc mô hình hiện tại (*Mô hình Recurrent với Attention*) để đánh giá điểm mạnh, hạn chế trong việc xử lý thông tin đa phương thức (hình ảnh và văn bản).
- Thực nghiệm trên bộ dữ liệu chuẩn *Chest X-rays [5]* để xác định các yếu tố ảnh hưởng đến độ chính xác của mô hình.

#### Kết quả dự kiến:

- Phát hiện các hướng đi tốt cho bài toán nhằm cải thiện hiệu suất và khai thác các mô hình mới để thay thế mô hình hiện tại nhằm nâng cao độ chính xác.
- Tìm ra cách cải thiện mô hình bằng việc ứng dụng các mô hình thay thế hoặc bổ sung như *Transformer* hoặc *Vision Transformer (ViT) [6]* để xử lý hình ảnh và văn bản hiệu quả hơn.
- Khám phá các yếu tố tác động như độ lớn của dữ liệu, phương pháp huấn luyện và kiến trúc mô hình.

### Nội dung 2: Nghiên cứu về các độ đo (metrics) cho việc đánh giá hiệu suất

Cần sử dụng các độ đo khác nhau để đánh giá chất lượng ngữ nghĩa (semantic quality) của các báo cáo y khoa tự động.

#### **Phương pháp thực hiện:**

- Xem xét và tìm hiểu các độ đo cho chất lượng ngữ nghĩa: các độ đo như BLEU, ROUGE, METEOR sẽ được đánh giá để kiểm tra độ chính xác ngữ nghĩa của các báo cáo tự động sinh ra từ mô hình.
- Nghiên cứu các độ đo chuyên biệt cho các bài toán y khoa, nhằm đánh giá tính chính xác của các thuật ngữ y khoa trong báo cáo.

#### **Kết quả dự kiến:**

- Kết hợp các độ đo tự động để xây dựng hệ thống đánh giá toàn diện cho mô hình.
- Đánh giá chính xác hơn về sự phù hợp của báo cáo với thực tế y khoa, giúp cải thiện độ tin cậy của mô hình.

### **Nội dung 3: Cải tiến mô hình với kiến trúc mới**

Để giải quyết các hạn chế của mô hình gốc, mô hình sẽ được thay thế hoặc bổ sung bằng các kiến trúc hiện đại hơn như *Vision Transformer (ViT)* cho việc xử lý đặc trưng hình ảnh và *BERT/BioBERT*,... cho việc xử lý văn bản.

#### **Phương pháp thực hiện:**

- *Xử lý hình ảnh:* Thay thế CNN bằng Vision Transformer (ViT) để khai thác tối đa các đặc trưng không gian trong hình ảnh, rất quan trọng trong phân tích hình ảnh y khoa.
- *Xử lý văn bản:* Sử dụng BioBERT hoặc BERT, thử nghiệm các mô hình tương tự khác để tạo sinh các báo cáo từ hình ảnh y khoa, tăng khả năng hiểu ngữ nghĩa.
- *Kết hợp thông tin đa phương thức:* Sử dụng các kỹ thuật như Fusion để đồng bộ hóa việc xử lý hình ảnh và văn bản, giúp mô hình học hỏi các đặc trưng đa phương thức hiệu quả.

#### **Kết quả dự kiến:**

- Mô hình mới sẽ sinh ra báo cáo chính xác hơn, với độ chính xác cao hơn so với mô hình gốc nhờ vào việc sử dụng các kiến trúc hiện đại.
- Khả năng xử lý thông tin đa phương thức sẽ được cải thiện, giúp mô hình hiểu và sinh báo cáo y khoa tốt hơn.

### **Nội dung 4: Kết hợp và chuẩn hóa các bộ dữ liệu**

Các bộ dữ liệu y khoa hiện tại còn hạn chế về số lượng và độ đa dạng. Việc kết hợp và chuẩn hóa các bộ dữ liệu lớn từ các nguồn khác nhau sẽ giúp mô hình học hỏi nhiều đặc trưng hơn, nâng cao hiệu suất và khả năng tổng quát hóa.

#### **Phương pháp thực hiện:**

- Kết hợp các bộ dữ liệu lớn như *MIMIC-CXR*, *ChestX-ray14*, *OpenI*, và *CheXpert*. Trong quá trình thử nghiệm, lựa chọn các bộ dữ liệu phù hợp nhất cho việc kết hợp.
- Chuẩn hóa dữ liệu về kích thước hình ảnh, định dạng văn bản và chú thích để đảm bảo tính tương thích giữa các bộ dữ liệu.

#### **Kết quả dự kiến:**

- Tạo ra một tập dữ liệu đa dạng và phong phú, giúp mô hình học nhiều đặc trưng khác nhau.
- Giảm thiểu overfitting khi huấn luyện với một tập dữ liệu lớn và đa dạng hơn.

### Nội dung 5: Huấn luyện và đánh giá mô hình đã đề xuất

Với mô hình cải tiến, thực hiện huấn luyện và đánh giá để kiểm chứng hiệu quả so với mô hình gốc.

#### Phương pháp thực hiện:

- Áp dụng mô hình mới với các kiến trúc đã đề xuất.
- Huấn luyện mô hình trên tập dữ liệu đã chuẩn bị.
- Tính toán các chỉ số như *BLEU*, *METEOR*, *ROUGE* và *KA* để đánh giá khả năng tổng quát của mô hình.

#### Kết quả dự kiến:

- Đánh giá hiệu quả mô hình mới so với mô hình gốc, từ đó đưa ra những nhận định về sự cải thiện trong độ chính xác và khả năng xử lý thông tin đa phương thức.

### KẾT QUẢ MONG ĐỢI

- Tăng cường khả năng hiểu và kết hợp thông tin từ hình ảnh và văn bản, nhờ sử dụng các kiến trúc hiện đại như *Vision Transformer (ViT)* và *BioBERT* (hoặc các mô hình tương tự khác), giúp mô hình cải tiến có khả năng tạo sinh báo cáo chính xác hơn.
- Một hệ thống sinh báo cáo y khoa tự động với độ chính xác cao, được cải thiện rõ rệt so với mô hình gốc.
- Tạo ra một tập dữ liệu lớn, đa dạng và chuẩn hóa từ các nguồn như *MIMIC-CXR*, *ChestX-ray14*, *CheXpert*, giúp giảm thiểu vấn đề overfitting và cải thiện khả năng tổng quát hóa.
- Xây dựng một mô hình có thể triển khai trong môi trường thực tế để hỗ trợ các bác sĩ lâm sàng trong việc phân tích hình ảnh y khoa và tạo báo cáo tự động.

### TÀI LIỆU THAM KHẢO

- [1]. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: CVPR, pp. 3337–3345 (2017).
- [2]. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* 9(8), 1735–1780 (1997).
- [3]. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G. R., & Huang, X.: Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In MICCAI 2018, LNCS 11070, Springer (2018). [https://doi.org/10.1007/978-3-030-00928-1\\_52](https://doi.org/10.1007/978-3-030-00928-1_52)
- [4]. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR, pp. 375–383 (2017).
- [5]. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In: CVPR, pp. 2497–2506 (2016).
- [6]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.: An image is worth 16x16 words: Transformers for image recognition at scale, (2020), doi:10.48550/arXiv.2010.11929