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

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

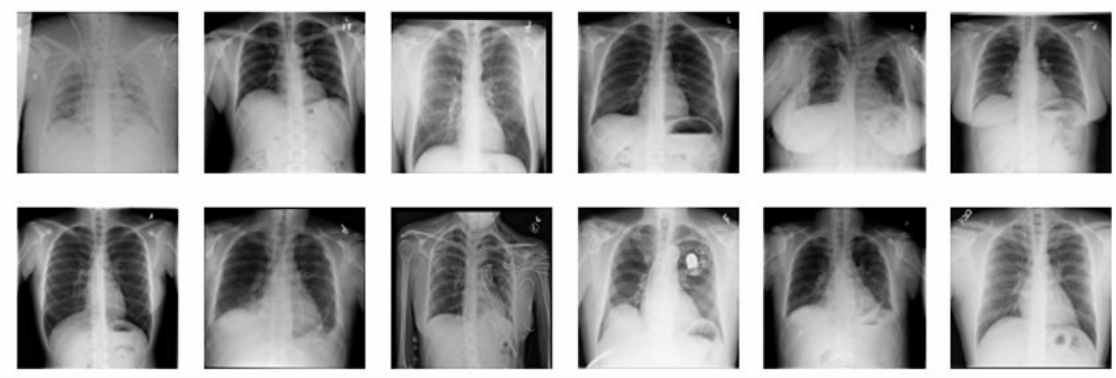
Thành viên thực hiện

- Họ và tên: Trần Minh Quân
- MSSV: 22521191
- Lớp: CS519.P11
- Link Github của nhóm: https://github.com/Be-Tap-Code/CS519.P11

Giới thiệu

- 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, cung cấp thông tin quan trọng về cấu trúc và chức năng của cơ thể.

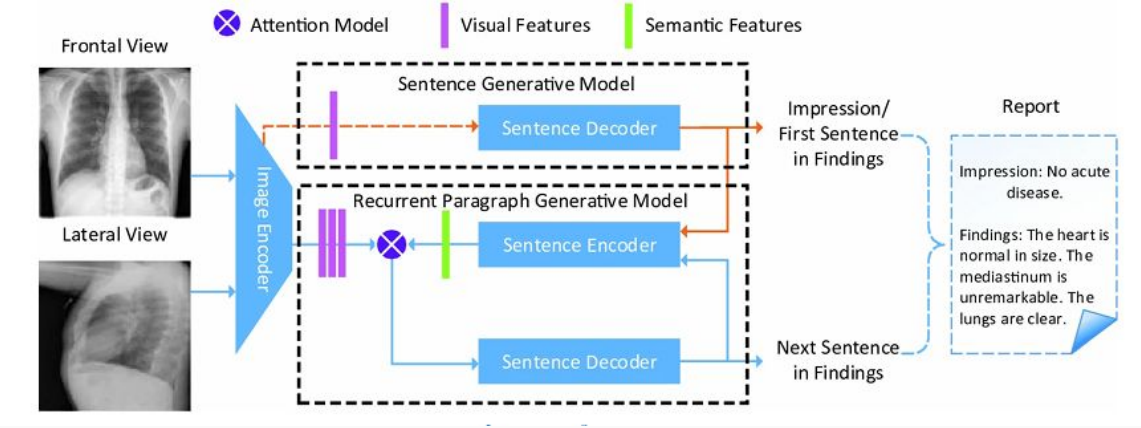
Những hạn chế, thách thức nào xuất hiện trong quá trình tự động tạo ra báo cáo X-quang?



Kết hợp đa modal (hình ảnh và ngữ nghĩa) với cơ chế Recurrent và Attention

Giới thiệu

Kiến trúc tổng quan từ các nghiên cứu trước, bao gồm 3 giai đoạn chính:



Giai đoạn 1: Xử lý hình ảnh đầu vào

Giai đoạn 2: Kết hợp đặc trưng ngữ nghĩa

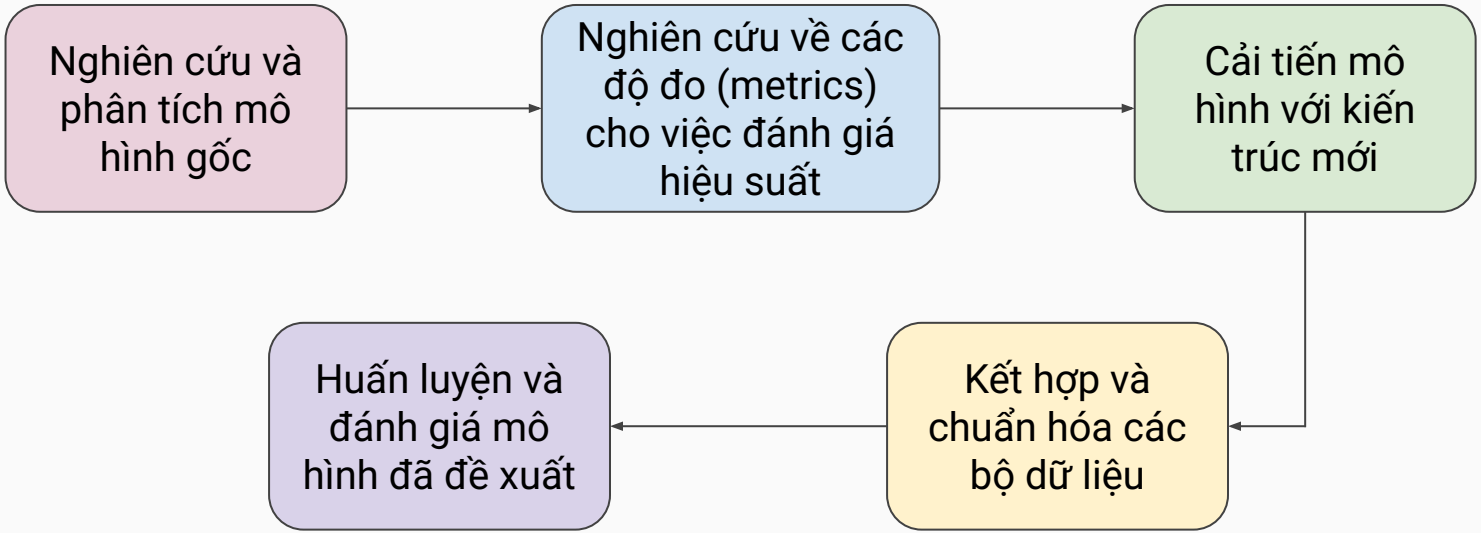
Giai đoạn 3: Tạo câu và đoạn văn (Report Generation)

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).

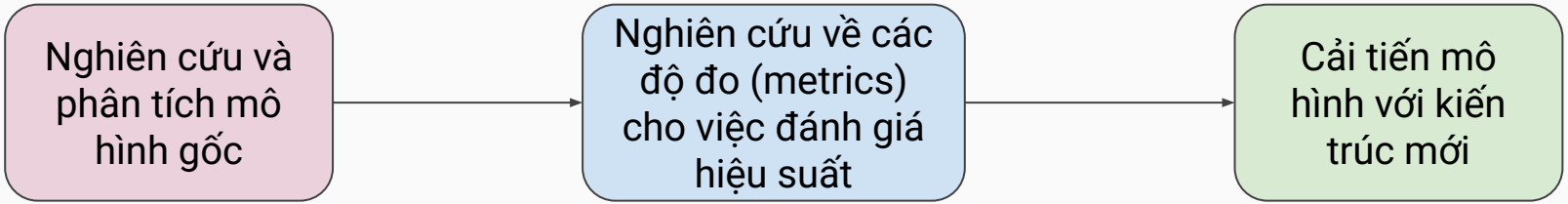
Mục tiêu

- **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 và Phương pháp



- 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*** để xác định các yếu tố ảnh hưởng đến độ chính xác của mô hình.

- 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.

- **Xử lý hình ảnh:** Thay thế CNN bằng Vision Transformer (ViT) để khai thác tối đa các đặc trưng không gian trong hình ảnh.
- **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**.

Nội dung và Phương pháp

- 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.

- 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

- 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.
- Xây dựng một **bộ công cụ đánh giá toàn diện** với các độ đo chính xác về ngữ nghĩa, tính nhất quán, và độ chính xác y khoa.
- 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. et al. (2018). Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In MICCAI 2018, LNCS 11070, Springer. 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