Ứng dụng Vision Transformer trong việc hỗ trợ phân tích và chẩn đoán ảnh chụp y sinh

Trần Hồng Đăng MSSV 22022646
Trần Kim Thành MSSV 22022532
Trịnh Minh Hiếu MSSV 22022536

Ngày 6 tháng 5 năm 2025

Abstract

Trí tuệ nhân tạo (AI) đang trở thành xu hướng chủ đạo trong nhiều lĩnh vực của xã hội, góp phần nâng cao hiệu suất công việc và cải thiện chất lượng cuộc sống. Hướng đến việc ứng dụng AI trong y tế, nhóm nghiên cứu đề xuất một mô hình Vision Transformer (ViT) có khả năng phát hiện các bệnh lý phổi thông qua phân tích ảnh X-quang ngực. Mô hình được tích hợp cơ chế diễn giải bằng bản đồ nhiệt thông qua phương pháp ReciproCAM, giúp minh bạch hóa quyết định của mô hình. Được huấn luyện trên bộ dữ liệu đa dạng, mô hình cho thấy độ chính xác cao và khả năng tổng quát tốt, qua đó hỗ trợ bác sĩ trong chẩn đoán nhanh chóng và hiệu quả, góp phần giảm tải cho hệ thống y tế.

1 Mở đầu.

Các phương pháp chụp, chiếu y sinh là một trong những kĩ thuật quan trọng và phổ biến nhất trong việc khám và điều trị bệnh. Từ tấm phim của bệnh nhân, bác sĩ sẽ có thể chẩn đoán chính xác và cụ thể những bất thường của nội quan cơ thể. Tại Việt Nam, ngày càng có nhiều bệnh nhân được chẩn đoán mắc các bệnh về phổi với nhiều nguyên nhân khác nhau. Tuy nhiên, với lực lượng y tế và cơ sở vật chất hiện tại, việc hỗ trợ cho tất cả bệnh nhân kịp thời đã trở thành một thử thách lớn, đặc biệt đối với các bệnh viện tuyến dưới. Thực trạng này đã phát sinh nhu cầu có một công cụ giúp hỗ trợ phân tích và chẩn đoán lâm sàng. Nhóm đã tiến hành nghiên cứu và xây dựng một mô hình trí tuệ nhân tạo có khả năng phân tích ảnh X-quang phổi, phát hiện bốn loại bệnh phổi phổ biến, đồng thời cung cấp khả năng diễn giải kết quả bằng bản đồ nhiệt để hỗ trợ bác sĩ trong quá trình đưa ra quyết định.

2 Công trình liên quan.

Class Activation Mapping: Là một mô hình Weakly-supervised object localization (WSOL), được xây dựng để học cách xác định vị trí đồ vật chỉ dựa trên nhãn dán của ảnh. Mô hình được xây dựng cho các mạng neuron CNN với lớp global average pooling. Kỹ thuật này cho phép các mô hình phân loại CNN có thể học và định vị vật thể mà

không cần sử dụng bounding box annotations. CAM làm nổi bật các vùng được CNN xác định trong quá trình phân loại.

Kiến trúc Transformer: Là mô hình học sâu được xây dựng cho việc xử lý ngôn ngữ tự nhiên. Mô hình gồm hai thành phần chính: encoder và decoder. Encoder có nhiệm vụ mã hóa đầu vào thành các biểu diễn ngữ nghĩa, còn decoder tạo ra đầu ra dựa vào các biểu diễn đó. Mỗi lớp trong encoder và decoder đều bao gồm các thành phần như multi-head self-attention, mạng feed-forward cộng với lớp chuẩn hóa và kết nối tắt. Cơ chế self-attention cho phép mô hình tập trung vào các từ liên quan trong câu bất kể vị trí, trong khi multi-head self-attention giúp học nhiều mối quan hệ ngữ nghĩa cùng lúc. Transformer còn sử dụng positional encoding để thêm thông tin về vị trí của từ trong chuỗi.

3 Phương pháp tiếp cận.

3.1 Ý tưởng thực hiện.

Nhóm đã lên ý tưởng một mô hình có thể thực tốt tác vụ về phân tích ảnh y tế, và một thuật toán có thể trích xuất vùng mà mô hình chú ý tới nhiều nhất cho mục đích diễn giải. Vision Transformer (ViT) là một trong những kiến trúc được sử dụng rộng rãi trong lĩnh vực phân tích ảnh do hiệu năng tiên tiến, nên nhóm đã quyết định áp dụng mô hình này làm xương sống của bài nghiên cứu. Về thuật toán giải thích, nhóm lựa chọn thuật toán ReciproCAM với khả năng tương thích tốt nhất đối với mô hình ViT.

Vì vậy, dự án sẽ bao gồm 3 thành phần chính:

- Huấn luyện: Sử dụng mô hình được tiền huấn luyện ViT để huấn luyện dữ liệu
- Sinh bản đồ nhiệt: Mô hình sau khi được huấn luyện sẽ được lưu lại để sử dụng cho ReciproCAM trong việc sinh bản đồ nhiệt.
- Kiểm thử: Kiểm tra lại trên các bộ dữ liệu khác nhau để đảm bảo mô hình có thể hoạt động tốt kể cả trên những dữ liệu mới và kiểm tra độ hiệu quả của ReciproCAM..

4 Kiến trúc và kĩ thuật.

4.1 Vision Transformer.

Kiến trúc Vision Transformer (ViT) là một trong những phương pháp tiên tiến hiện nay trong lĩnh vực phân tích hình ảnh và nhận diện vật thể. Khác với các kiến trúc CNN truyền thống vốn sử dụng các kernel tích chập cục bộ, ViT chia hình ảnh đầu vào thành các patch cố định, sau đó ánh xạ từng patch thành vector và đưa vào mô hình Transformer với cơ chế self-attention gồm 12 lớp. Kiến trúc ViT-base mà nhóm sử dụng có 12 khối Transformer và vector embedding 768 chiều cho mỗi patch. Việc này cho phép ViT học được các mối liên hệ dài hạn và phụ thuộc toàn cục giữa các vùng ảnh.

Trên nhiều tập dữ liệu benchmark như ImageNet, ViT đã chứng minh khả năng vượt trội so với các kiến trúc CNN sâu như ResNet khi được huấn luyện trên tập dữ liệu đủ lớn. Đặc biệt, trong lĩnh vực ảnh y tế, nơi thông tin hình ảnh mang tính chi tiết cao và phụ thuộc không gian phức tạp, cơ chế attention của ViT giúp nắm bắt tốt hơn mối liên kết giữa các vùng tổn thương và cấu trúc tổng thể, từ đó cải thiện độ chính xác trong các tác vụ như phân đoạn, phát hiện bất thường, và chẩn đoán tự động.

Với những ưu điểm rõ rệt, nghiên cứu của nhóm sẽ sử dụng mô hình ViT-base của Hug-

ging Face đã được tiền huấn luyện trên tập dữ liệu ImageNet-21k với patch có kích thước 16x16.

4.2 ReciproCAM.

ReciproCAM là một phương pháp giải thích trực quan cho các mô hình học sâu trong lĩnh vực thị giác máy tính, với một nghiên cứu dành riêng cho Vision Transformer (ViT). Khác với các kỹ thuật dựa trên attention hay gradient truyền thống, ReciproCAM cho ViT khai thác cơ chế thử nghiệm cục bộ (perturbation) để đánh giá mức độ đóng góp của từng vùng ảnh đối với kết quả dự đoán.

Cụ thể, ảnh đầu vào được chia thành các patch và đưa qua mô hình ViT và ReciproCAM sẽ trích xuất các bản đồ đặc trưng từ đầu ra của lớp LayerNorm ở khối transformer encoder cuối cùng nơi chứa thông tin đặc trưng đã tổng hợp từ toàn bộ ảnh. Sau đó, thực hiện quá trình phân tích như sau: với ảnh đầu vào, phương pháp lần lượt giữ lại từng patch cùng với 8 patch lân cận (áp dụng 3x3 Gaussian kernel), đồng thời che hoàn toàn các patch còn lại trong ảnh. Ảnh bị che được đưa qua cùng một mô hình để thu lại điểm logit đầu ra tương ứng. Mức chênh lệch trong điểm logit của lớp được dự đoán so với những vị trí bị che khác sẽ là mức độ ảnh hưởng của patch đó đến phân loại cuối cùng. Vì vậy, vùng nào có giá trị logits của nhãn đúng cao nhất chính là vùng mô hình chú ý đến.

Sau khi lặp lại quy trình này cho tất cả các patch trong ảnh, ReciproCAM tạo bản đồ nhiệt (saliency map) bằng giá trị logits của từng vùng ảnh, trong đó các vùng màu đỏ biểu thị những patch có logits cao nhất, đóng góp vào kết quả đầu ra. Bản đồ này phản ánh trực tiếp mối quan hệ nhân quả giữa vùng ảnh và đầu ra của mô hình, thay vì chỉ dựa trên trọng số attention nội tại.

Phương pháp này không chỉ hiệu quả hơn trong việc tạo bản đồ trực quan, mà còn hoạt động tốt hơn trên các mô hình không dùng gradient rõ ràng như ViT. Vì vậy, trong ứng dụng với ảnh y tế, ReciproCAM sẽ giúp bác sĩ hiểu được lý do mô hình đưa ra quyết định, từ đó tăng tính tin cậy và khả năng diễn giải trong thực tế lâm sàng.

5 Thí nghiệm.

5.1 Bộ dữ liệu huấn luyện.

Nhóm sử dụng bộ dữ liệu "COVID-19 Radiography Database" của Tawsifur Rahman, Muhammad Chowdhury và Amith Khandakar - một bộ dữ liệu phổ biến trên Kaggle bao gồm 21,150 hình ảnh X-quang ngực (CXR - Chest X-Ray) có các bệnh lý phổi. Bộ dữ liệu có điểm đánh giá sử dụng trên Kaggle là 10.0, được ủng hộ 1073 lần và đạt giải nhất bộ dữ liệu Covid-19 do người dùng bình chọn:

- Covid: Nhóm này chứa hình ảnh X-quang ngực của bệnh nhân mắc bệnh COVID-19. (3616 ảnh).
- Lung Opacity: Ảnh X-quang của các bệnh lý gây đục phổi khác (6012 ảnh).
- Viral Pneumonia: Nhóm này đại diện cho các trường hợp viêm phổi do virus khác ngoài COVID-19. Hình ảnh X-quang có thể cho thấy các vùng mờ kính hoặc thâm nhiễm lan tỏa, nhưng thường ít nghiêm trọng hơn so với COVID-19.(1345 ảnh).
- Normal: Nhóm này bao gồm hình ảnh X-quang của những người khỏe mạnh, không có dấu hiệu bệnh phổi. (10,192 ảnh).

5.2 Quy trình huấn luyện.

Nhóm đã tiến hành fine-tune mô hình bằng tập dữ liệu "COVID-19 Radiography Database" trên Kaggle với tốc độ học 2e-05, phương pháp tối ưu ADAM, hàm mất mát cross-entropy, áp dụng cross-validation và early stopping với giá trị patience = 5 trong 15 epochs. Sử dụng mô hình đã được tiền huấn luyện giúp nhóm tiết kiệm thời gian phát triển, đồng thời tận dụng được khả năng trích xuất đặc trưng đã học với bộ ImageNet-21k.

Bộ dữ liệu huấn luyện đã cung cấp ảnh định dạng .png với kích thước 224x224 phù hợp cho ViT, nên chỉ cần tạo feature extractor có sẵn của mô hình pre-train và sử dụng nó để tạo data loader. Feature extractor bao gồm các thao tác chỉnh kích thước, crop trung tâm, chuyển đổi thành RGB, chuyển đổi thành tensor và chuẩn hóa với mean=[0.5, 0.5, 0.5], std=[0.5, 0.5, 0.5].

5.3 Bộ dữ liệu kiểm thử.

Bộ dữ liệu kiểm thử được tổng hợp từ 7 bộ dữ liệu khác nhau trên Kaggle, bộ bao gồm 4800 ảnh với 1200 COVID, 1200 Normal, 1200 Viral Pneumonia và 1200 Lung Opacity từ các dataset:

- Covid-19 Image Dataset
- Pneumonia X-Ray Images
- Image Classification - Covid-19 X-Ray
- Lung Area Specific COVID-19 Xray Dataset
- COVID-19 Digital X-rays Forgery Dataset
- COVID-19 SARS MERS X-ray Images Dataset.
- covid_normal_viral_opacity_v2

Việc lấy mẫu ảnh một cách ngẫu nhiên từ 7 bộ dữ liệu và đồng đều về số lượng ảnh giữa các lớp được nhóm chú trọng để tạo ra kết quả đánh giá khách quan nhất về hiệu suất mô hình.

5.4 Thước đo đánh giá.

Để đánh giá hiệu suất của mô hình, nhóm đã áp dụng một số thước đo phổ biến, đồng thời ghi lại độ mất mát trên cả hai tập huấn luyện và tập đánh giá trong quá trình fine-tune.

- Accuracy: Tỷ lệ dự đoán đúng trên tổng số mẫu. Phản ánh hiệu suất tổng thể của mô hình.
- Recall: Khả năng mô hình phát hiện đúng các mẫu thuộc một lớp (TP / (TP + FN)).
- Precision: Tỷ lệ mẫu đúng trong số các mẫu được dự đoán là thuộc lớp đó (TP / (TP + FP)).
- F1 score: Trung bình điều hòa giữa precision và recall. Giá trị cao cho thấy mô hình cân bằng tốt giữa hai yếu tố.
- AUC-ROC: Diện tích dưới đường cong ROC. Giá trị càng cao (gần 1) cho thấy mô hình phân biệt giữa các lớp càng tốt.

Về thuật toán nâng cao tính diễn giải của mô hình, vì không có chuyên gia về y tế trong nhóm, nên tính lâm sàng của bản đồ nhiệt rất khó kiểm chứng, mà chỉ có thể xem thuật toán có chú ý vào vùng hai lá phổi hay không.

6 Kết quả thí nghiệm.

6.1 Kết quả kiểm thử.

Kết quả của việc fine-tune mô hình đã được pre-train trên bộ Image-21k rất đáng chú ý. Mô hình có xu hướng overfit trong những epoch cuối, và nhờ có early-stopping, mô hình tốt nhất với mất mát trên tập đánh giá nhỏ nhất đã được lưu lại ở epoch 7. Với thời gian khoảng 9 phút 1 epoch, mô hình tốt nhất chỉ cần hơn 1 tiếng huấn luyện (epoch 7) với độ chính xác 99% trên tập huấn luyện, 98% trên tập kiểm tra và đánh giá (bộ huấn luyện) và 95% trên tập kiểm thử (tổng hợp 7 bộ ngoài) (Hình 1).

Kết quả kiểm tra sử dụng bộ dữ liệu kiểm thử cho thấy độ hiệu quả của mô hình ở mức tốt và toàn diện với:

- **Accuracy**: 96.02%
- **Recall**: 0.960
- **Precision**: 0.961
- **F1 score**: 0.960
- **AUC-ROC**: 0.996

6.2 Phân tích kết quả.

	Model	Accuracy	Precision	Recall	F1-Score
COVID-19 Radiography Database	ResNet101	98%	0.980	0.983	0.980
	InceptionV3	97%	0.966	0.966	0.970
	MobileNetV2	98.08%	0.976	0.976	0.9766
	Xception	97%	0.973	0.973	0.976
	InstaCovNet-19	99.08%	0.99	0.99	0.99
	VGG19	95%	0.94	0.96	0.95
	Covid-Net	93.3%	0.92	0.91	0.93
	ViT của nhóm	96.02%	0.961	0.960	0.960

Bảng 1: So sánh các mô hình trên COVID-19 Radiography Database.

Kết quả thống kê cho thấy mô hình của nhóm đạt được kết quả tốt, có độ chính xác cao trong việc phân loại ảnh, đồng thời duy trì sự cân bằng tốt giữa khả năng phát hiện chính xác các trường hợp dương tính (recall) và giảm thiểu tỷ lệ báo động giả (precision). Đặc biệt, chỉ số AUC-ROC đạt 0.996 phản ánh khả năng phân biệt xuất sắc giữa bốn lớp trong mô hình (Hình 2).

Mặc dù kết quả của nhóm chưa vượt trội so với một số mô hình khác trên cùng tập dữ liệu, hướng tiếp cận của nhóm lại mang tính mới mẻ khi là một trong những nghiên cứu đầu tiên ứng dụng Vision Transformer (ViT) kết hợp với ReciproCAM trong bài toán chẩn đoán ảnh X-quang phổi. Các nghiên cứu trước đó chủ yếu dựa trên các kiến trúc CNN truyền thống và chưa giải quyết triệt để bài toán thiếu tính diễn giải của mô hình. Nhằm tăng cường khả năng giải thích, nhóm đã áp dụng thuật toán ReciproCAM để xác định các vùng ảnh có ảnh hưởng mạnh đến quyết định của mô hình. Kết quả cho thấy mô hình ViT không chỉ có độ chính xác cao mà còn mang lại khả năng giải thích đáng tin cậy, mở ra hướng mới cho ứng dụng trong chẩn đoán y tế (Hình 3).

7 Tổng kết

Mô hình của nhóm đã chứng minh khả năng phân loại chính xác các bệnh phổi với các chỉ số kiểm thử ấn tượng, đồng thời thể hiện năng lực tổng quát hóa vượt trội trên dữ liệu mới. Việc tích hợp ReciproCAM giúp cải thiện tính diễn giải của mô hình, từ đó tăng cường sự tin tưởng của bác sĩ và bệnh nhân đối với các quyết định chẩn đoán. Giải pháp đề xuất không chỉ nâng cao độ chính xác mà còn đảm bảo tính minh bạch và đạo đức trong ứng dụng AI y tế, góp phần giảm tải cho hệ thống y tế và đưa dịch vụ khám chữa bệnh công nghệ cao đến gần hơn với cộng đồng.

A Phụ lục.

Code: https://github.com/MinHius/ViT-CAM-for-Biomedical-Image-Analysis

A.1 Phân công công việc.

- Trần Hồng Đăng: Thuyết trình, viết báo cáo (phần thí nghiệm, tổng hợp kết quả các nghiên cứu cùng lĩnh vực), nghiên cứu và lựa chọn các bộ dữ liệu, rà soát kết quả nghiên cứu và nội dung báo cáo.

- Trần Kim Thành: Làm PowerPoint, viết báo cáo (phần công trình liên quan, kết quả thí nghiệm, tài liệu, kết luận), định hướng nghiên cứu của nhóm, rà soát kết quả và nội dung báo cáo.

- Trịnh Minh Hiếu: Huấn luyện và kiểm thử mô hình, cài đặt thuật toán ReciproCAM, viết báo cáo (abstract, mở đầu, kiến trúc và kỹ thuật, phụ lục), thảo luận định hướng nghiên cứu và xử lý kỹ thuật.

A.2 Những khó khăn còn tồn tại.

Ngoài những kết quả ấn tượng đã thu được, nghiên cứu của nhóm vẫn còn một số điều có thể khắc phục trong nghiên cứu sau này:

- Mô hình của nhóm cần được tinh chỉnh và huấn luyện thêm để đạt hiệu suất vượt trội hơn.
- Thuật toán ReciproCAM hoạt động tốt với ViT và cho ra kết quả rõ ràng, nhưng tính chính xác không ổn định vì trong nhiều trường hợp kết quả đầu ra không mang ý nghĩa lâm sàng.
- Bộ dữ liệu kiểm thử dù giữa các lớp có số lượng ảnh xấp xỉ, chưa hoàn toàn theo phân phối đều giữa số lượng nguồn của các lớp do những giới hạn về thu thập dữ liệu.

A.3 Định hướng tương lai.

Trong tương lai, những vấn đề nêu trong mục trên cần được khắc phục bằng những biện pháp như tinh chỉnh tham số để mô hình hội tụ sâu hơn và đạt kết quả state-of-the-art. Thuật toán diễn giải vốn không đảm bảo tính chất lâm sàng, nhưng có thể thử lại sau khi huấn luyện mô hình với tham số được tinh chỉnh và bộ dữ liệu học lớn hơn. Khi đó, có thể mô hình sẽ nhận biết được vùng phổi là vùng duy nhất mang tính quyết định.

Ngoài ra, dự án của nhóm vẫn còn rất nhiều hướng phát triển, trong đó thực tiễn nhất là mở rộng bộ dữ liệu huấn luyện để không chỉ có thể chẩn đoán nhiều loại bệnh, mà cả những tổn thương vật lý và dị vật. Hơn nữa, mô hình có tiềm năng điều chỉnh để nhận đầu vào là ảnh CT hoặc MRI để nâng cao hiệu quả đánh giá và tính linh hoạt của công trình. Trong tương lai, mô hình có thể được kết hợp cùng các module khác như

mô hình ngôn ngữ để vừa tạo ra báo cáo y tế, vừa giúp giao tiếp với bác sĩ, bệnh nhân để cải thiện khả năng tương tác và hợp tác.

Tài liệu

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[2] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023, p. 7433–7442. [Online]. Available: http://dx.doi.org/10.1109/CVPR52729.2023.00718

[3] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," 2020. [Online]. Available: https://arxiv.org/abs/2003.09871

[4] A. Gupta, Anjum, S. Gupta, and R. Katarya, "Instacovnet-19: A deep learning classification model for the detection of covid-19 patients using chest x-ray," *Applied Soft Computing*, vol. 99, p. 106859, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494620307973

[5] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "Medvit: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, p. 106791, May 2023. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2023.106791

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2015. [Online]. Available: https://arxiv.org/abs/1512.04150

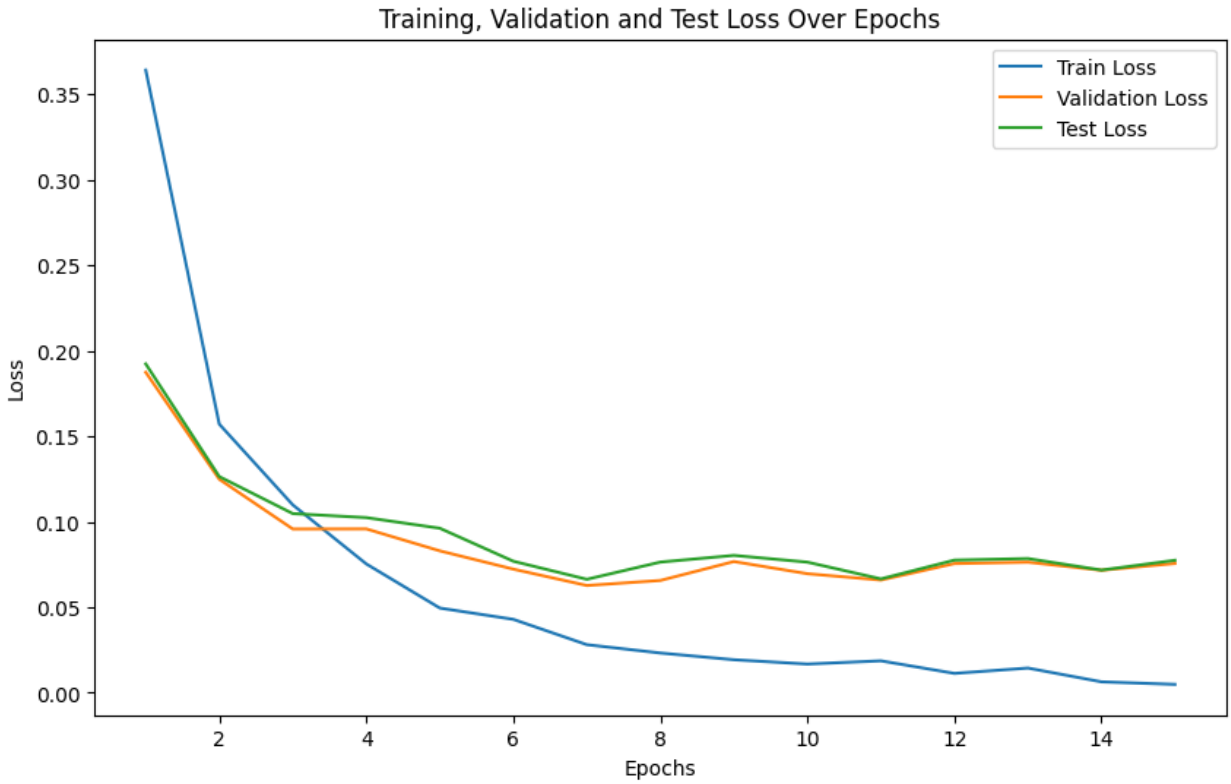
[7] S.-Y. Byun and W. Lee, "Reciprocam: Fast gradient-free visual explanations for convolutional neural networks," 2023. [Online]. Available: https://arxiv.org/abs/2209.14074

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

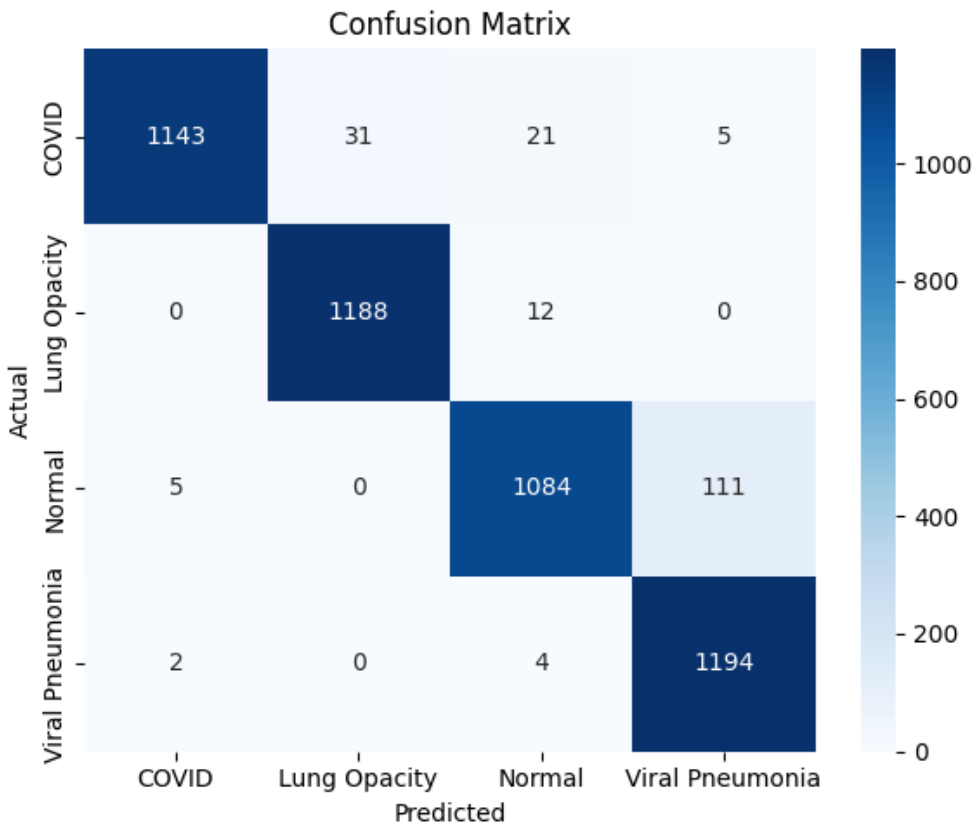
[9] E. M. F. El Houby, "Covid-19 detection from chest x-ray images using transfer learning," *Scientific Reports*, vol. 14, no. 1, p. 11639, May 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-61693-0

[10] S.-Y. Byun and W. Lee, "Vit-reciprocam: Gradient and attention-free visual explanations for vision transformer," 2023. [Online]. Available: https://arxiv.org/abs/2310.02588

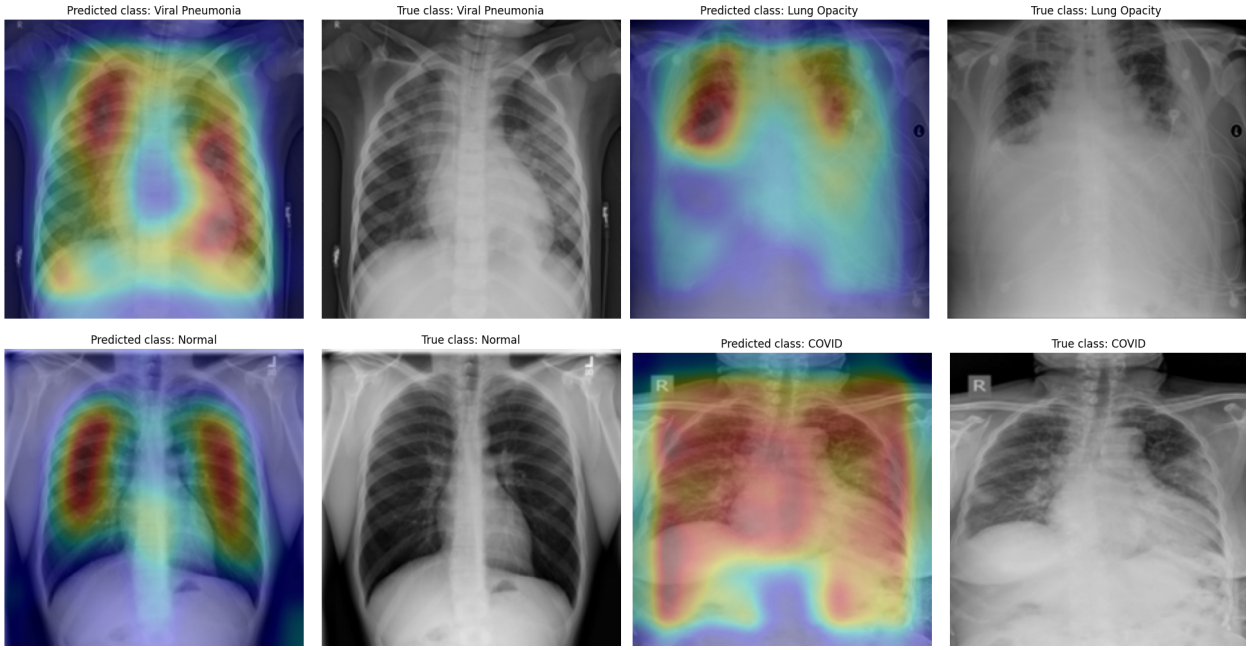
[1] [2] [3] [4] [5] [6] [7] [8] [2] [9] [10]



Hình 1: Kết quả huấn luyện.



Hình 2: Kết quả kiểm thử.



Hình 3: Đầu ra của mô hình.