ĐẠI HỌC FPT – VIỆN ĐÀO TẠO SAU ĐẠI HỌC

Chương trình Thạc sĩ Kỹ thuật Phần mềm

Nghiên cứu So sánh CNN, ResNet và Vision Transformer cho Phân loại Đa nhãn Bệnh lý X-quang Ngực

Môn học: DLE501 – Deep Learning

Giảng viên: TS. [Tên giảng viên]

Nhóm 1:

[Thành viên 1]

[Thành viên 2]

[Thành viên 3]

Thành phố Hồ Chí Minh – 2026

Tóm tắt nghiên cứu

Tổng quan

Bài báo "A Comparative Study of CNN, ResNet, and Vision Transformers for Multi-Classification of Chest Diseases" (arXiv:2406.00237) so sánh hiệu năng của ba nhóm kiến trúc học sâu trong bài toán phân loại đa nhãn bệnh lý trên ảnh X-quang ngực sử dụng bộ dữ liệu NIH ChestX-ray14.
Các mô hình được nghiên cứu: CNN baseline, ResNet-34, ViT-v1 (từ đầu), ViT-v2 (cải tiến), và ViT-ResNet (pretrained).

Kết quả thực nghiệm chính

Mô hình	Params	Test AUC	Test Acc	Epochs	Ghi chú
CNN	∼95.6M	0.5777	–	10	Overfitting nghiêm trọng
ResNet-34	∼21.3M	0.4462	–	10	Underfitting
ViT-v1	9.0M	0.5854	91.33%	10	Huấn luyện từ đầu
ViT-v2	9.0M	0.6303	89.67%	9	SGD + Early stopping
ViT-ResNet	85.8M	0.6694	87.00%	10	Pretrained ImageNet
ViT Final	9.0M	0.7225	92.91%	10	Full dataset 112K

Kết quả trên tập con nhỏ (60 ảnh train) cho thấy ViT-ResNet pretrained đạt AUC cao nhất (0.6694). Khi chạy trên toàn bộ 112,120 ảnh với patient-level split, ViT đạt AUC 0.7225.

MỤC LỤC

DANH SÁCH HÌNH

DANH SÁCH BẢNG

1. Giới thiệu

1.1 Bối cảnh lâm sàng

Bệnh lý phổi và tim–phổi là nguyên nhân hàng đầu gây tử vong toàn cầu. X-quang ngực là phương tiện chẩn đoán phổ biến nhất nhờ chi phí thấp, tốc độ nhanh, và khả năng phát hiện đa dạng bệnh lý.

1.2 Thách thức

- Thiếu hụt bác sĩ chẩn đoán hình ảnh, đặc biệt tại vùng sâu vùng xa.
- Biến thiên giữa các bác sĩ trong đánh giá cùng một ảnh.
- Bản chất đa nhãn: một ảnh có thể mang nhiều bệnh đồng thời.
- Dấu hiệu bệnh tinh vi, nhiều tổn thương giai đoạn sớm rất khó phát hiện.

1.3 Mục tiêu nghiên cứu

Bài báo tập trung so sánh CNN, ResNet-34, và Vision Transformer trên bộ dữ liệu NIH ChestX-ray14 với 112,120 ảnh và 14 bệnh lý. Các câu hỏi nghiên cứu:

RQ1 Hiệu năng CNN, ResNet và ViT khác nhau thế nào?

RQ2 Transfer learning cải thiện bao nhiêu so với huấn luyện từ đầu?

RQ3 Yếu tố nào quyết định hiệu năng: kiến trúc hay dữ liệu?

2. Bộ dữ liệu NIH ChestX-ray14

2.1 Thông tin tổng quan

Bảng 2.1: Thông tin bộ dữ liệu NIH ChestX-ray14

Thuộc tính	Giá trị
Tổng số ảnh	112,120
Số bệnh nhân	30,805
Số lớp bệnh	14 + No Finding = 15
Kích thước gốc	1024×1024 pixels
Format	PNG (grayscale $\rightarrow$ RGB)

2.2 15 lớp bệnh lý

Bảng 2.2: 15 Classes trong NIH ChestX-ray14

ID	Tên bệnh	Tỷ lệ (%)	Mô tả
0	Cardiomegaly	2.48	Tim to
1	Emphysema	2.24	Khí phế thũng
2	Effusion	11.88	Tràn dịch màng phổi
3	Hernia	0.20	Thoát vị
4	Nodule	5.65	Nốt phổi
5	Pneumothorax	4.73	Tràn khí màng phổi
6	Atelectasis	10.31	Xẹp phổi
7	Pleural_Thickening	3.02	Dày màng phổi
8	Mass	5.16	Khối u
9	Edema	2.05	Phù phổi
10	Consolidation	4.16	Đông đặc phổi
11	Infiltration	17.74	Thâm nhiễm
12	Fibrosis	1.50	Xơ phổi
13	Pneumonia	1.28	Viêm phổi
14	No Finding	53.84	Bình thường

2.3 Mất cân bằng lớp

Vấn đề Class Imbalance

- "No Finding" chiếm 53.84% — hơn một nửa dataset.
- "Hernia" chỉ chiếm 0.20% — hiếm nhất.
- Tỷ lệ chênh lệch cao nhất/thấp nhất = 53.84 / 0.20 = 269 lần.

2.4 Multi-label Classification

Mỗi ảnh có thể mang nhiều nhãn bệnh đồng thời. Do đó sử dụng Sigmoid (thay vì Softmax) và BCEWithLogitsLoss.

2.5 Chia tập dữ liệu

Thí nghiệm 1 (Small-scale): 60 train / 20 val / 20 test ảnh.

Thí nghiệm 2 (Full-scale): 112,120 ảnh chia theo Patient ID:

- Train: 78,614 ảnh (21,563 bệnh nhân)
- Val: 11,212 ảnh (3,081 bệnh nhân)
- Test: 22,294 ảnh (6,161 bệnh nhân)

3. Convolutional Neural Network (CNN)

3.1 Kiến trúc

CNN baseline gồm 2 lớp tích chập (32 và 64 filters, kernel 3×3), mỗi lớp đi kèm ReLU và MaxPool 2×2, sau đó Flatten, FC 512 nút, Dropout 0.5, và lớp đầu ra 15 nút.

3.2 Phân tích tham số

Bảng 3.1: Tham số CNN baseline

Lớp	Số tham số	Tỷ lệ
Conv layers	∼19,400	0.02%
FC layers	∼95,560,000	99.98%
Tổng	∼95.6M	100%

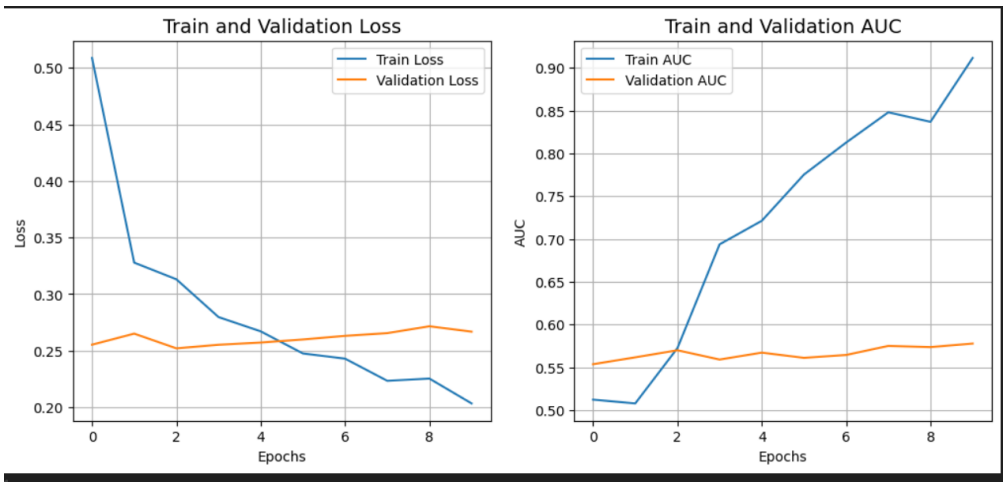
Vấn đề

99% tham số nằm ở FC layer ⇒ overfitting nghiêm trọng. Train AUC = 0.9116 nhưng Test AUC chỉ đạt 0.5777.

3.3 Kết quả thực nghiệm

Bảng 3.2: Kết quả CNN (small-scale, 60 ảnh train)

Metric	Train	Val	Test
Loss	0.2034	0.2668	–
AUC	0.9116	0.5777	0.5777



Hình 3.1: Diễn biến huấn luyện CNN: overfitting rõ rệt

4. Residual Network (ResNet-34)

4.1 Ý tưởng cốt lõi: Skip Connection

ResNet giải quyết degradation problem bằng residual learning:

$$y = F(x) + x$$

Gradient luôn có thành phần "1" đi qua skip connection, tránh vanishing gradient.

4.2 Kiến trúc ResNet-34

Bảng 4.1: Kiến trúc ResNet-34

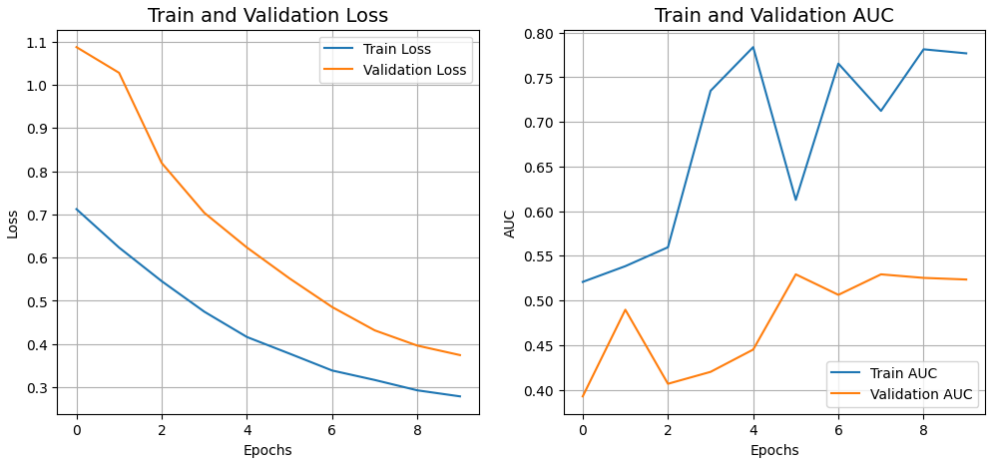
Stage	Output	Channels	Blocks	Stride
Conv1	112×112	64	–	2
MaxPool	56×56	64	–	2
Layer1	56×56	64	3	1
Layer2	28×28	128	4	2
Layer3	14×14	256	6	2
Layer4	7×7	512	3	2
AvgPool + FC	–	15	–	–

Tổng: ∼21.3M tham số (ít hơn nhiều so với CNN 95.6M).

4.3 Kết quả thực nghiệm

Bảng 4.2: Kết quả ResNet-34 (small-scale, 60 ảnh train)

Metric	Train	Val	Test
Loss	0.3065	0.3742	–
AUC	0.7342	0.4462	0.4462



Hình 4.1: Diễn biến huấn luyện ResNet-34

Phân tích

ResNet-34 huấn luyện từ đầu trên dữ liệu nhỏ cho kết quả kém (AUC 0.4462). Kiến trúc sâu cần pretrained weights hoặc dữ liệu lớn để phát huy hiệu quả.

5. Vision Transformer (ViT)

5.1 Kiến trúc tổng quan

ViT chuyển bài toán ảnh thành bài toán chuỗi:

1. Chia ảnh 224×224 thành patches $32 \times 32 \Rightarrow 49$ patches.
2. Linear projection sang embedding dimension 64.
3. Thêm positional embedding + [CLS] token.
4. Đưa qua 8 Transformer Encoder blocks (4 attention heads).
5. MLP Head: [CLS] output $\rightarrow$ 15 classes (sigmoid).

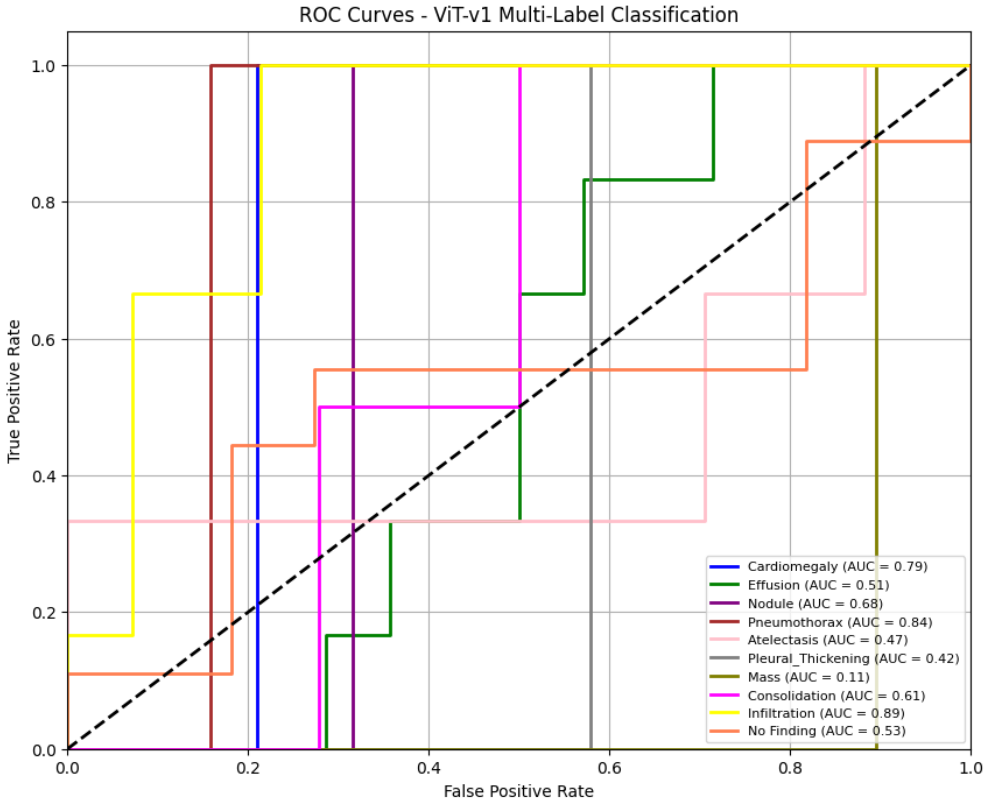
Tổng: 9,005,839 tham số.

5.2 ViT-v1: Huấn luyện từ đầu

Cấu hình: Adam optimizer, lr $= 10^{-4}$, weight decay $= 10^{-6}$, 10 epochs.

Bảng 5.1: Kết quả ViT-v1 (small-scale)

Metric	Train	Val	Test
Loss	0.2569	0.2717	0.2534
Accuracy	90.56%	90.00%	91.33%
AUC	0.5883	0.5868	0.5854



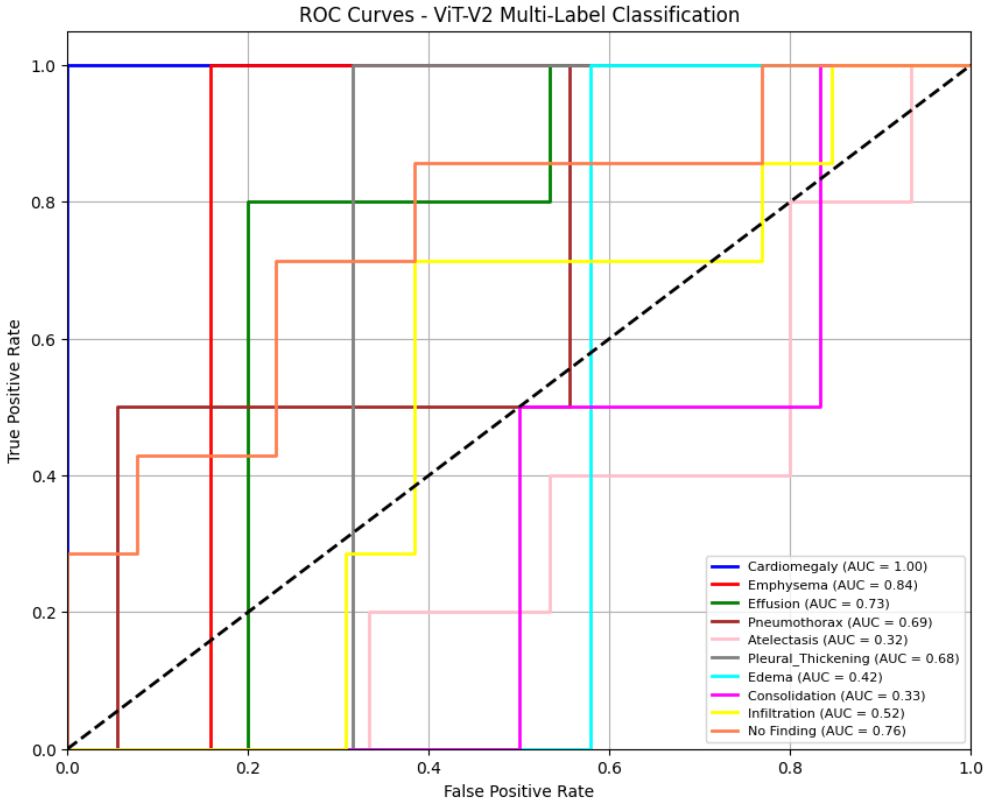
Hình 5.1: ROC curves ViT-v1 theo từng lớp bệnh

5.3 ViT-v2: Cải tiến với SGD và Early Stopping

Thay đổi so với v1: SGD optimizer, weight decay = 10^{-5}, early stopping patience = 3.

Bảng 5.2: Kết quả ViT-v2 (small-scale)

Metric	Train	Val	Test
Loss	0.2657	0.2413	0.2749
Accuracy	89.78%	91.67%	89.67%
AUC	0.5630	0.5947	0.6303



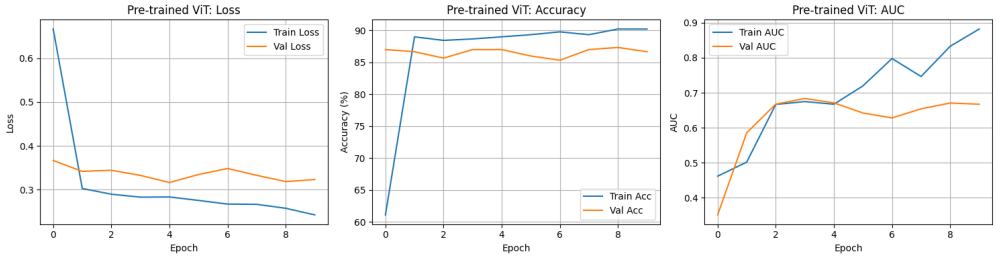
Hình 5.2: ROC curves ViT-v2

5.4 ViT-ResNet: Pretrained (timm vit_base_patch16_224)

Sử dụng vit_base_patch16_224 pretrained trên ImageNet. Tổng: 85,810,191 tham số.

Bảng 5.3: Kết quả ViT-ResNet pretrained (small-scale)

Metric	Train	Val	Test
Loss	0.2425	0.3232	0.3768
Accuracy	90.22%	86.67%	87.00%
AUC	0.8820	0.6673	0.6694



Hình 5.3: Diễn biến huấn luyện ViT-ResNet pretrained

6. Thí nghiệm Full-scale: ViT trên 112,120 ảnh

6.1 Cấu hình

- Dataset: 112,120 ảnh, chia theo Patient ID
- Split: Train 78,614 / Val 11,212 / Test 22,294
- Patients: 21,563 / 3,081 / 6,161
- Model: ViT (patch 32, proj_dim 64, 4 heads, 8 layers)
- Epochs: 10, Batch size: 32

6.2 Kết quả

Bảng 6.1: Kết quả ViT Full-scale (112K ảnh)

Metric	Train	Val	Test
Loss	0.1985	0.1967	0.2001
Accuracy	92.93%	93.01%	92.91%
AUC (Macro)	0.7195	0.7264	0.7225

6.3 AUC theo từng lớp bệnh

Bảng 6.2: Per-class AUC trên tập Test (ViT Full-scale)

Bệnh	AUC	Bệnh	AUC
Edema	0.8422	Pleural_Thick.	0.6997
Cardiomegaly	0.7996	Fibrosis	0.6977
Effusion	0.7880	Mass	0.6762
Consolidation	0.7615	Pneumonia	0.6710
Pneumothorax	0.7540	Infiltration	0.6614
Hernia	0.7460	Nodule	0.5747
Emphysema	0.7375		
Atelectasis	0.7170	Macro Avg	0.7225
No Finding	0.7114		

7. So sánh và Phân tích

7.1 Bảng so sánh tổng hợp

Bảng 7.1: So sánh tất cả các mô hình

Model	Data	Params	Test AUC	Test Acc	Test Loss	Epochs
CNN	60	95.6M	0.5777	–	–	10
ResNet-34	60	21.3M	0.4462	–	–	10
ViT-v1	60	9.0M	0.5854	91.33%	0.2534	10
ViT-v2	60	9.0M	0.6303	89.67%	0.2749	9
ViT-ResNet	60	85.8M	0.6694	87.00%	0.3768	10
ViT Final	112K	9.0M	0.7225	92.91%	0.2001	10

7.2 Phân tích chính

1. Dữ liệu quyết định: Cùng kiến trúc ViT, AUC tăng từ 0.5854 (60 ảnh) lên 0.7225 (112K ảnh) — cải thiện 23.4%.
2. Transfer learning hiệu quả: ViT-ResNet pretrained đạt AUC 0.6694 chỉ với 60 ảnh, vượt tất cả mô hình from-scratch trên dữ liệu nhỏ.
3. CNN không phù hợp cho dữ liệu ít: 99% tham số ở FC layer dẫn đến overfitting (train AUC 0.91 vs test 0.58).
4. ResNet cần dữ liệu lớn: Huấn luyện từ đầu trên 60 ảnh cho AUC thấp nhất (0.4462).
5. ViT-v2 cải thiện v1: SGD + early stopping giúp tăng AUC từ 0.5854 lên 0.6303.

8. Kết luận và Hướng phát triển

8.1 Kết luận

1. Vision Transformer đạt hiệu năng tốt nhất trên full dataset (AUC 0.7225, Acc 92.91%).
2. Transfer learning là yếu tố then chốt khi dữ liệu hạn chế.
3. Quy mô dữ liệu quan trọng hơn kiến trúc mô hình.
4. CNN đơn giản bị overfitting do cấu trúc FC quá lớn.

8.2 Hướng phát triển

- Sử dụng Focal Loss / Asymmetric Loss để xử lý class imbalance.
- Thử nghiệm Swin Transformer cho multi-scale features.
- Ensemble nhiều mô hình để tăng hiệu năng.
- Tăng epochs và sử dụng learning rate scheduling.
- Đánh giá trên CheXpert và MIMIC-CXR để kiểm chứng tính tổng quát.

9. Cấu hình Huấn luyện

Bảng 9.1: Cấu hình huấn luyện chung

Tham số	Giá trị
Image size	224×224
Batch size	32 (16 cho ViT-ResNet)
Num classes	15
Loss function	BCEWithLogitsLoss
Learning rate	1×10^{-4}
Weight decay	1×10^{-6} (10^{-5} cho ViT-v2)
Optimizer	AdamW (SGD cho ViT-v2)
Epochs	10
GPU	NVIDIA GeForce RTX 3060 Laptop (6.4 GB)
CUDA	12.6
Framework	PyTorch 2.x

Tài liệu tham khảo

[1] Jain, M. et al. (2024). A Comparative Study of CNN, ResNet, and Vision Transformers for Multi-Classification of Chest Diseases. arXiv:2406.00237.

[2] Wang, X. et al. (2017). ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks. CVPR.

[3] Dosovitskiy, A. et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.

[4] He, K. et al. (2016). Deep Residual Learning for Image Recognition. CVPR.

[5] Rajpurkar, P. et al. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225.

[6] Lin, T.-Y. et al. (2017). Focal Loss for Dense Object Detection. ICCV.

[7] Ridnik, T. et al. (2021). Asymmetric Loss For Multi-Label Classification. ICCV.

[8] Vaswani, A. et al. (2017). Attention Is All You Need. NeurIPS.

[9] Liu, Z. et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV.

[10] Irvin, J. et al. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. AAAI.