

ĐÁNH GIÁ HIỆU NĂNG CỦA CÁC MÔ HÌNH HQ U-NET TRONG BÀI TOÁN PHÂN ĐOẠN HÌNH ẢNH Y TẾ

Võ Văn Thành^{1*}, Lê Sỹ Toàn²

¹ Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế

² K45F, Khoa công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế

*Email: vvthanh@husc.edu.vn

Ngày nhận bài: 8/12/2024; ngày hoàn thành phản biện: 18/12/2024; ngày duyệt đăng: 20/12/2024

TÓM TẮT

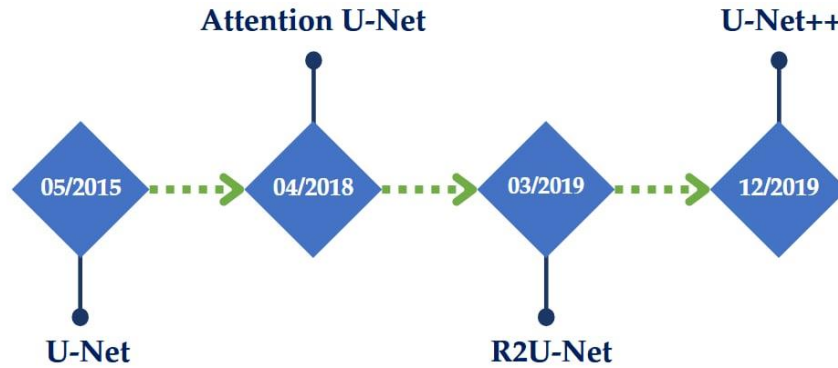
Phân đoạn hình ảnh y tế là một ứng dụng quan trọng trong lĩnh vực học sâu, đóng vai trò then chốt trong hỗ trợ chẩn đoán, điều trị và phân tích các bệnh lý trong y học. Trong nghiên cứu này, chúng tôi sẽ giới thiệu kiến trúc và so sánh hiệu năng của các mô hình họ U-Net bao gồm U-Net, Attention U-Net, U-Net++ và R2U-Net trong phân đoạn hình ảnh trên hai tập dữ liệu hình ảnh y tế: tổn thương da và X-quang lồng ngực, dựa vào một số độ đo phổ biến như DSC và IoU. Kết quả cho thấy R2U-Net đạt hiệu năng phân đoạn tốt nhất trên cả hai tập dữ liệu (DSC > 0.91 trên tập ảnh tổn thương da và DSC > 0.98 trên tập ảnh X-quang lồng ngực) trong khi các mô hình còn lại phụ thuộc vào đặc điểm từng loại dữ liệu. Nghiên cứu cung cấp cơ sở cho việc lựa chọn các mô hình U-Net tối ưu phù hợp, thúc đẩy việc chẩn đoán hình ảnh chính xác hơn. Toàn bộ mã nguồn và thông tin về tập dữ liệu của nghiên cứu được công khai ở <https://github.com/LeSyToan2003/Medical-Segmentation.git>.

Từ khóa: Phân đoạn hình ảnh y tế, Học sâu, U-Net, Các biến thể U-Net, Mạng nơ-ron tích chập.

1. MỞ ĐẦU

Bài toán phân đoạn hình ảnh được giới thiệu lần đầu tiên vào năm 2012 bởi D. Ciresan và các cộng sự [1], nhằm xác định các vùng đặc trưng trong hình ảnh. Trong nhiều năm trở lại đây, việc áp dụng học sâu vào phân đoạn hình ảnh đóng góp thành công lớn trong lĩnh vực y tế, điển hình với kiến trúc U-Net. Nhiều kiến trúc cải tiến từ mô hình U-Net như Attention U-Net, U-Net++, R2U-Net đã nâng cao hiệu năng phân đoạn trên nhiều tập dữ liệu về hình ảnh y tế [5-11]. Tuy nhiên, vẫn còn thiếu các nghiên cứu thực nghiệm trực tiếp so sánh hiệu quả giữa các biến thể U-Net trên nhiều dạng ảnh y tế khác nhau. Việc đánh giá định lượng toàn diện là cần thiết để hỗ trợ lựa chọn mô hình tối ưu cho từng loại hình ảnh có độ nhiễu và cấu trúc phức tạp.

Mạng tích chập đầy đủ (Fully Convolutional Networks - FCN) từng là nền tảng cho các bài toán nhận dạng và phân đoạn hình ảnh [2-4], nhưng gặp thách thức về dữ liệu huấn luyện lớn và độ phức tạp tính toán trong quá trình xử lý [15]. Mô hình U-Net ra đời với mục đích giải quyết các thử thách đó, sử dụng dạng kiến trúc Mã hóa - Giải mã (Encoder - Decoder) kết hợp đường kết nối tắt (Skip Connection) [5]. Theo sự tiến hóa trong lĩnh vực học sâu, các biến thể được cải tiến từ U-Net chứng tỏ được tiềm năng phát triển của dạng mô hình này.



Hình 1: Các mô hình họ U-Net

Mục tiêu chính của bài nghiên cứu là giới thiệu kiến trúc và so sánh hiệu năng các mô hình họ U-Net trên hai tập dữ liệu ảnh y tế gồm tổn thương da và X-quang lồng ngực thông qua các độ đo đánh giá. Bằng cách đánh giá kết quả thực nghiệm của từng mô hình để cung cấp cơ sở trong việc lựa chọn kiến trúc phù hợp cho các bài toán phân đoạn hình ảnh y tế với độ nhiễu và cấu trúc phức tạp của hình ảnh. Trong nghiên cứu này, chúng tôi không áp dụng kỹ thuật tăng dữ liệu (augmentation), do dữ liệu đầu vào có tính đa dạng về hình thái và số lượng mẫu lớn, đảm bảo đại diện tốt cho các biến thể ảnh cần phân đoạn.

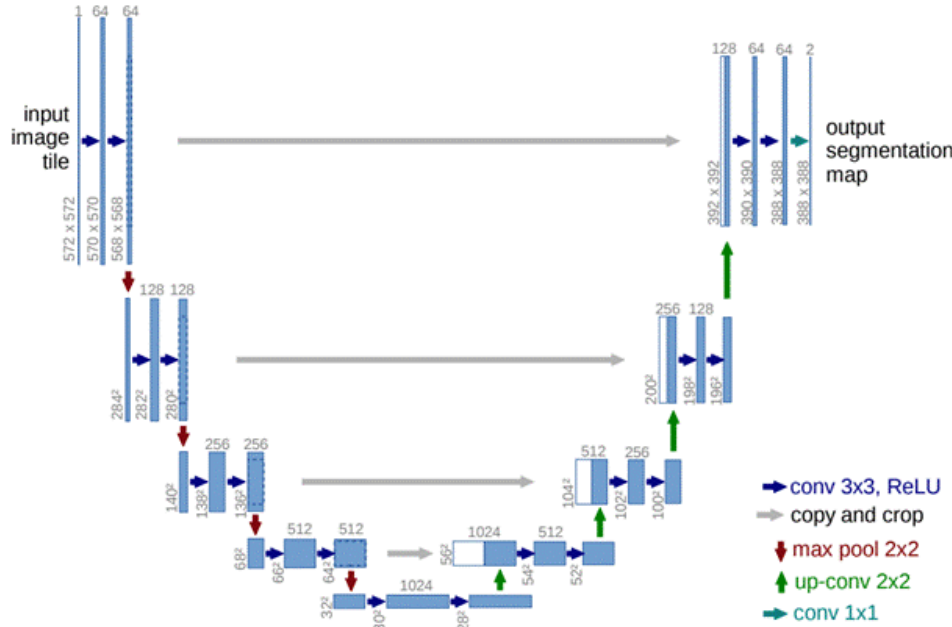
2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Các kiến trúc mô hình

2.1.1. Mô hình U-Net đầu tiên

Vào năm 2015, Ronneberger và các cộng sự đã đề xuất kiến trúc U-Net và đã đạt được chiến thắng trong cuộc thi phân đoạn ảnh y tế ISBI Cell Tracking Challenge [5]. U-Net gồm nhánh mã hóa (Encoder) và nhánh giải mã (Decoder) mỗi nhánh có bốn khối tích chập. Nhánh mã hóa sử dụng hai lớp tích chập và một lớp max-pooling để trích xuất thông tin thông qua giảm dần kích thước không gian và tăng chiều sâu đặc trưng. Nhánh giải mã với một lớp tích chập chuyển vị và hai lớp tích chập liên tiếp nhằm khôi phục lại kích thước không gian ảnh. Các đường kết nối tắt truyền thông tin từ nhánh mã hóa sang nhánh giải mã làm giảm mất mát đặc trưng chi tiết và cải thiện độ chính xác phân

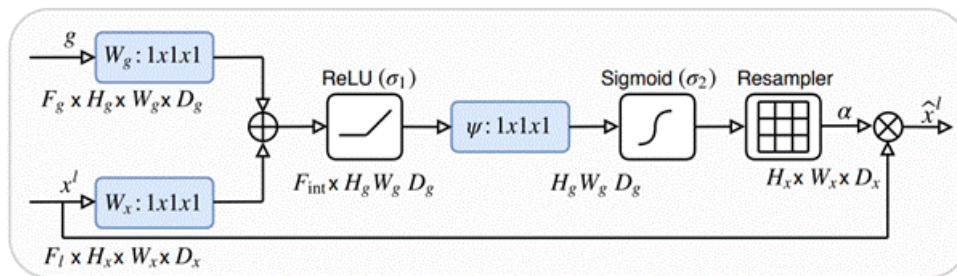
đoạn. Lớp tích chập cuối giúp ánh xạ thông tin thành bản đồ phân đoạn cuối cùng (segmentation map).



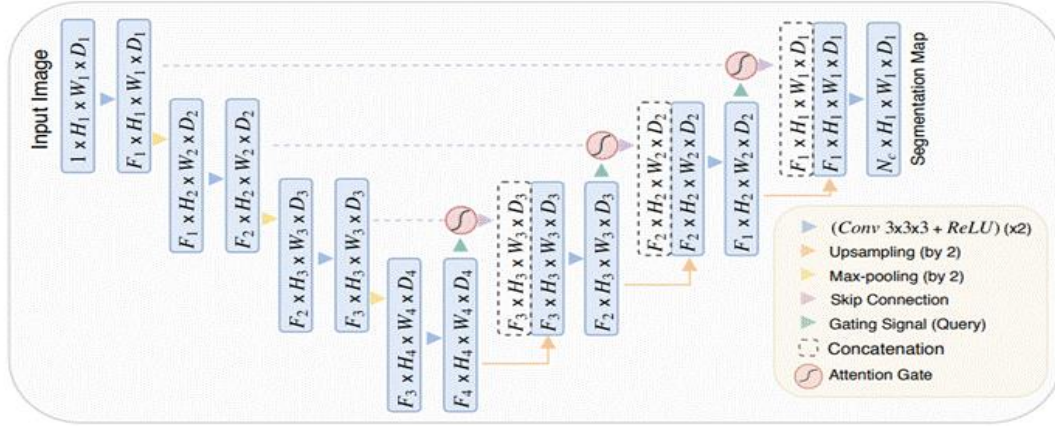
Hình 2. Kiến trúc mô hình U-Net [5], gồm nhánh mã hóa (Encoder) và nhánh giải mã (Decoder)

2.1.2 Mô hình Attention U-Net và kiến trúc Cổng chú ý

Được O. Oktay và các cộng sự đề xuất năm 2018 [6], Attention U-Net tích hợp kiến trúc Cổng chú ý (Attention Gate - AG) [12-14] vào các lớp kết nối của U-Net (Concatenation Layer). Thông qua kiến trúc AG, thông tin quan trọng được trích xuất và loại bỏ phần nhiễu trong đường kết nối tắt. Quá trình tính toán này được thực hiện trước khi đưa vào lớp kết nối nhằm chỉ kết hợp các thông tin có ý nghĩa.



Hình 3: Kiến trúc Cổng chú ý (AG) [6]

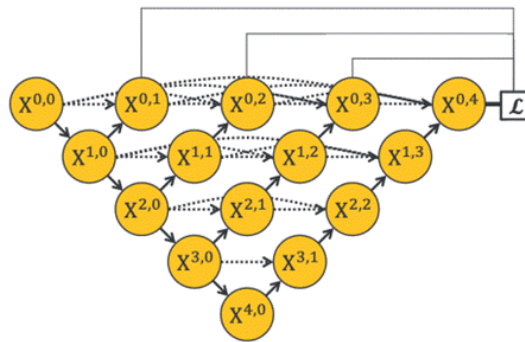


Hình 4: Kiến trúc mô hình Attention U-Net [6], giữ lại kiến trúc U-Net và bổ sung thêm AG tại các đường kết nối tắt

Theo R. Azad và các cộng sự, AG giúp loại bỏ các thông tin không liên quan trong bản đồ đặc trưng, làm nổi bật các đặc trưng quan trọng, yếu tố cần thiết trong các bài toán phân đoạn hình ảnh [11]. Mặc dù AG tăng số lượng tham số và tốc độ xử lý nhưng lại giúp cải thiện độ chính xác phân đoạn [10, 11].

2.1.3. Mô hình U-Net++ và sự mở rộng các đường kết nối tắt

Nghiên cứu của X. Li và các cộng sự khẳng định vai trò của các đường kết nối tắt trong mô hình U-Net giúp cải thiện phân đoạn bằng cách bảo toàn thông tin giữa hai nhánh mã hóa và giải mã [7]. Tuy nhiên ở mô hình U-Net chưa tối ưu hóa việc kết hợp đặc trưng. Năm 2019 Zhou và các cộng sự đã giới thiệu mô hình U-Net++ với sự mở rộng các đường kết nối tắt [10]. Khác với mô hình U-Net khi mỗi khối giải mã được kết nối với một khối mã hóa tương ứng, ở kiến trúc U-Net++, chúng đi qua các khối tích chập dày đặc, số lượng khối này phụ thuộc vào độ sâu của vị trí.



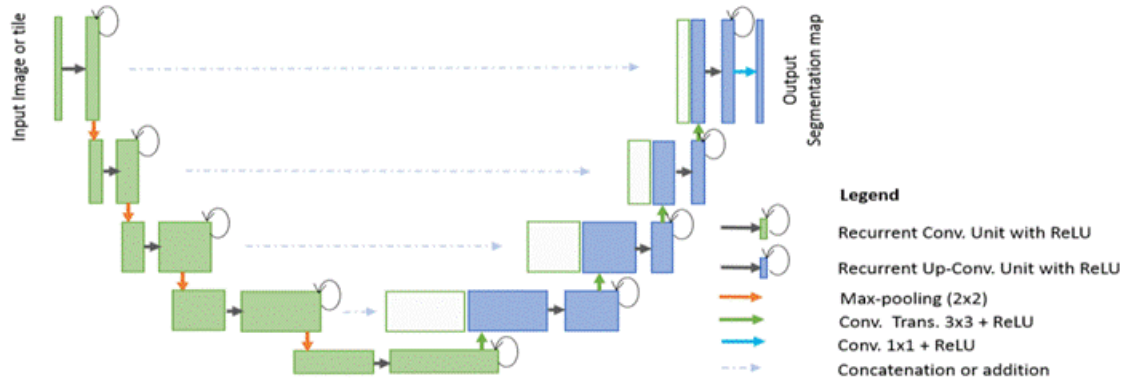
Hình 5: Kiến trúc mô hình U-Net++ [8], mở rộng các đường kết nối tắt qua các khối tích chập dày đặc với số khối phụ thuộc vào độ sâu mạng.

Việc bổ sung thêm các khối trên đường kết nối, đồng thời tăng số lượng đường kết nối giữa chúng, tạo nên một mạng dày đặc các khối tích chập giúp kiến trúc U-Net++

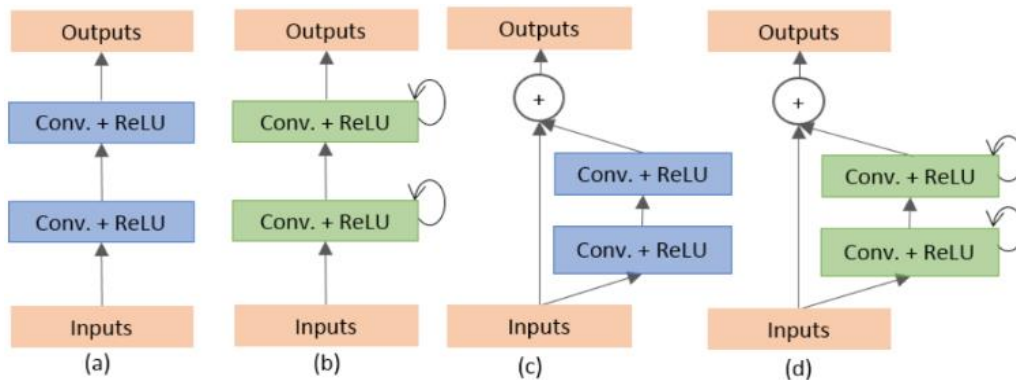
học được cách kết hợp đặc trưng tốt hơn, thay vì chỉ đơn giản kết nối trực tiếp chúng [8]. Nghiên cứu của R. Azad chỉ ra rằng việc mở rộng này đồng nghĩa việc đánh đổi độ phức tạp tính toán với việc trao đổi khả năng xử lý hiệu quả của mô hình, từ đó U-Net++ thích hợp với các bài toán phân đoạn hình ảnh phức tạp [11].

2.1.4. Mô hình R2U-Net và kiến trúc Khối phần dư

Được đề xuất bởi M. Z. Alom và các cộng sự [9], Recurrent U-Net (RU-Net) và Recurrent Residual U-Net (R2U-Net) được cải tiến từ U-Net, với khả năng giải quyết hiệu quả các bài toán phân đoạn hình ảnh y tế, kết hợp cả hai đặc điểm của mạng nơ-ron tích chập hồi tiếp (Recurrent Convolutional Neural Networks - RCNN) [3] và khối phần dư (Residual Block) [4] vào mô hình U-Net truyền thống. Các lớp tích chập ban đầu trở thành các lớp tích chập hồi tiếp (Recurrent Convolutional Layer - RCL), có khả năng liên kết các vùng ảnh với nhau nhờ cơ chế đặc biệt của nó, thay vì chỉ xử lý từng vùng ảnh nhỏ như kiến trúc ban đầu.



Hình 6: Kiến trúc mô hình RU-Net [9]



Hình 7: Các kiến trúc khối tích chập (a) Khối tích chập truyền thống (Convolutional Block) (b) Khối tích chập hồi tiếp (Recurrent Convolutional Block) (c) Khối tích chập phần dư (Residual Convolutional Block) (d) Khối tích chập hồi tiếp - phần dư (Recurrent Residual Convolutional Block) [9]

Hình 7 mô tả các dạng kiến trúc phát triển từ khối tích chập ban đầu, bao gồm các sự kết hợp của hồi tiếp (recurrent) và phần dư (residual). Khối tích chập phần dư sử dụng

dùng đường kết nối tắt để kết hợp thông tin đầu vào với đầu ra của các lớp tích chập, có khả năng làm giảm mất mát thông tin và hạn chế hiện tượng vanishing gradient [11]. Bảng 1 dưới đây thể hiện sự so sánh giữa các mô hình họ U-net.

Bảng 1. So sánh đặc điểm chính của các mô hình họ U-Net

Mô hình	Kiến trúc chính	Ưu điểm chính	Hạn chế
U-Net	Encoder–Decoder, Skip Connection đơn giản	Dễ huấn luyện, hiệu quả với dữ liệu đơn giản	Dễ mất thông tin ở tầng sâu, không xử lý tốt ảnh nhiễu
Attention U-Net	Thêm Attention Gate vào các skip connection	Giảm nhiễu, làm nổi bật vùng quan trọng	Tăng số tham số, có thể loại bỏ nhầm thông tin cần thiết
U-Net++	Skip connection mở rộng, tích chập dày đặc	Học đặc trưng sâu, hiệu quả với ảnh có cấu trúc phức tạp	Phức tạp tính toán, dễ bị overfitting
R2U-Net	Recurrent + Residual blocks trong encoder–decoder	Học tốt ngữ cảnh sâu, giữ thông tin, giảm vanishing gradient	Tốn tài nguyên, thời gian huấn luyện dài hơn

2.2. Huấn luyện và đánh giá mô hình

Hàm mất mát Binary Cross-Entropy được áp dụng với bài toán phân đoạn một đối tượng (phân lớp nhị phân), cụ thể việc tính toán ở công thức sau:

$$L_{BCE} = -\frac{1}{n} \sum ((y_i \cdot \log(\hat{y}_i)) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \quad (1)$$

Trong đó, n là số lượng mẫu, y_i là giá trị thực và $\hat{y}_i$ là giá trị dự đoán.

Các độ đo phổ biến được sử dụng để đánh giá khả năng phân đoạn của mô hình được tính toán dựa trên các giá trị: TP (True Positive) là số lượng điểm ảnh lớp dương được dự đoán đúng, TN (True Negative) là số lượng điểm ảnh lớp âm được dự đoán đúng, FP (False Positive) là số lượng điểm ảnh lớp dương được dự đoán sai và FN (False Negative) là số lượng điểm ảnh lớp âm được dự đoán sai. Độ đo Dice Score (DSC) được sử dụng để đánh giá mức độ trùng khớp giữa mặt nạ thực tế (ground-truth mask) và mặt nạ dự đoán (prediction mask) trên tập kiểm thử (testset). Giá trị Dice Score càng tiến về 1, mô hình càng cho thấy khả năng phân đoạn hiệu quả:

$$Dice\ Score = \frac{2TP}{2TP + FP + FN} \quad (2)$$

Ngoài Dice Score, việc đánh giá khả năng phân đoạn của mô hình trên tập kiểm thử được thực hiện dựa trên độ đo Intersection over Union (IoU):

$$IoU = \frac{TP}{TP + FP + FN} \quad (3)$$

Các độ đo phổ biến trong bài toán phân lớp cũng được sử dụng để đánh giá các mô hình, bao gồm Accuracy, Precision, Recall và F1-Score:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

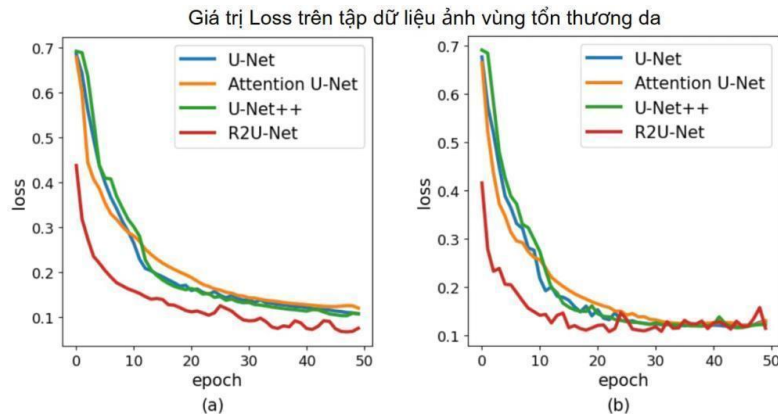
$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (7)$$

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Phân đoạn vùng tổn thương da

Chúng tôi sử dụng tập dữ liệu bao gồm 2594 hình ảnh về tổn thương da, được tổng hợp từ các loại ung thư da hoặc tổn thương da phổ biến (được chia theo tỷ lệ 85% cho tập huấn luyện và 15% cho tập kiểm thử) với kích thước 128x128x3, mô hình được huấn luyện với batch size 128, sử dụng optimizer Adam với learning_rate 1e-4. Kết quả so sánh thông qua các độ đo được thể hiện tại bảng 1 thông qua quá trình huấn luyện.



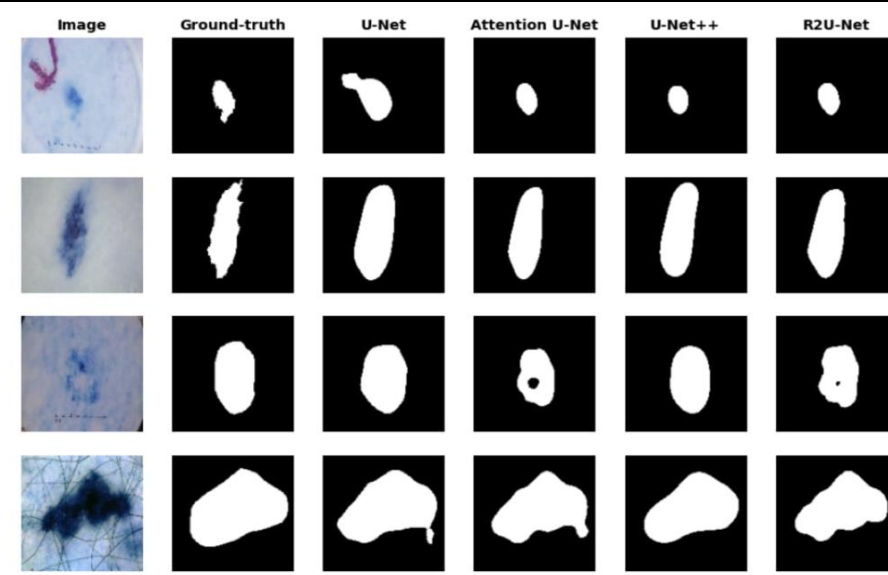
Hình 8: Giá trị hàm mất mát (Loss) trong quá trình huấn luyện trên tập dữ liệu ảnh vùng tổn thương da: (a) trên tập huấn luyện, (b) trên tập thẩm định

Bảng 2. Tham số, thời gian huấn luyện và số epoch của các mô hình cho tập tổn thương da

Mô hình	Tham số huấn luyện	Thời gian huấn luyện	Epoch
U-Net	2,158,705	5 phút	50
Attention U-Net	2,185,917	5 phút	50
U-Net++	2,478,853	4 phút	50
R2U-Net	6,373,201	8 phút	50

Bảng 3. Mô tả kết quả phân đoạn trên tập kiểm thử ảnh vùng tổn thương da

Mô hình	Loss	DSC	IoU	Accuracy	Precision	Recall	F1-Score
U-Net	0.1678	0.8659	0.835	0.9319	0.7537	0.9068	0.7911
Attention U-Net	0.1199	0.9031	0.8773	0.9506	0.8791	0.8672	0.8494
U-Net++	0.1017	0.9145	0.8889	0.9583	0.8482	0.9066	0.8596
R2U-Net	0.1027	0.9161	0.8911	0.9581	0.8702	0.9145	0.8746



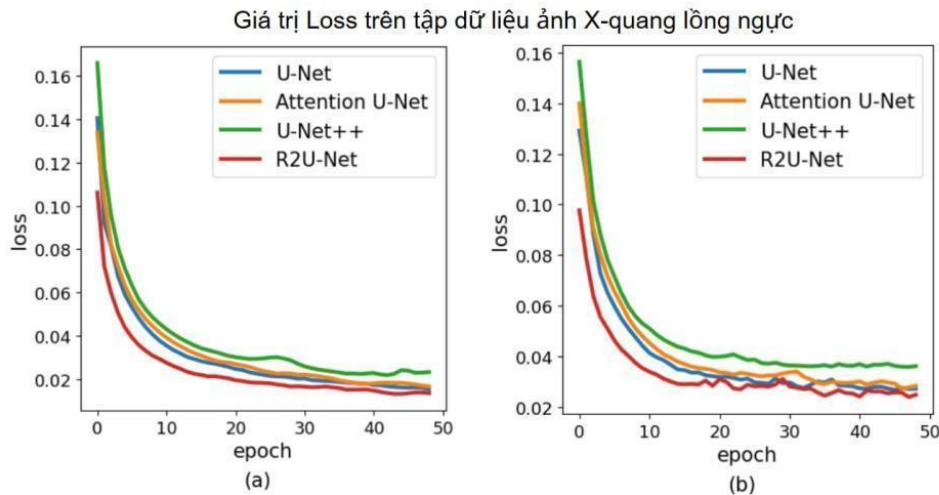
Hình 9: Một số kết quả phân đoạn vùng tổn thương da của các mô hình

Các hình ảnh tổn thương da có đặc điểm phức tạp, nhiều yếu tố nhiễu do đa dạng về hình dạng cấu trúc và màu sắc, biên không đều gây khó phân biệt các vùng cần được phân đoạn. U-Net bị ảnh hưởng bởi nhiễu của dữ liệu do thiếu cơ chế tập trung và học ngữ cảnh sâu, mô hình dễ phân đoạn sai vùng da lành thành vùng tổn thương (Precision thấp nhất đạt 0.75). Attention U-Net có kết quả tốt hơn U-Net do loại bỏ phần thông tin nhiễu và giảm sự mất mát thông tin thông qua kiến trúc AG (Precision đạt 0.87) nhưng bỏ sót một số chi tiết biên (Recall đạt 0.86). U-Net++ học chi tiết sâu hơn nhờ khối tích chập dày đặc, phù hợp với cấu trúc phức tạp nhưng nhiễu vẫn ảnh hưởng đến

mô hình (IoU đạt 0.88), nhưng nhiễu vẫn ảnh hưởng (Precision đạt 0.84). R2U-Net kết hợp cả hai khả năng hạn chế bỏ sót thông tin, giảm nhiễu thông qua Residual Block và học các chi tiết biên quan trọng nhờ Recurrent Convolutional Layer, từ đó hiệu năng phân đoạn của biến thể này vượt trội hơn (kết quả cao nhất IoU đạt 0.89 và DSC đạt 0.91) so với các mô hình còn lại, đặc biệt ở hai độ đo chuyên dụng để đánh giá khả năng phân đoạn là DSC và IoU.

3.2. Phân đoạn vùng phổi

Tập dữ liệu gồm 21165 hình ảnh X-quang lồng ngực, được tổng hợp từ 4 loại chính: ảnh của bệnh nhân nhiễm Covid-19, mắc bệnh viêm phổi, nhiễm mờ trắng phổi do các nguyên nhân khác và không mắc bệnh, (được chia theo tỷ lệ 90% cho tập huấn luyện và 10% cho tập kiểm thử) với kích thước 128x128x1, mô hình huấn luyện qua 50 epoch, batchsize 128, sử dụng optimizer Adam với learning_rate 2e-4. Kết quả đánh giá hiệu năng của từng mô hình trên tập kiểm thử được thể hiện ở bảng 2.



Hình 10: Giá trị hàm mất mát (Loss) trong quá trình huấn luyện trên tập dữ liệu ảnh X-quang lồng ngực: (a) trên tập huấn luyện, (b) trên tập thẩm định

Bảng 4. Tham số, thời gian huấn luyện và số epoch của các mô hình cho tập X-quang lồng ngực

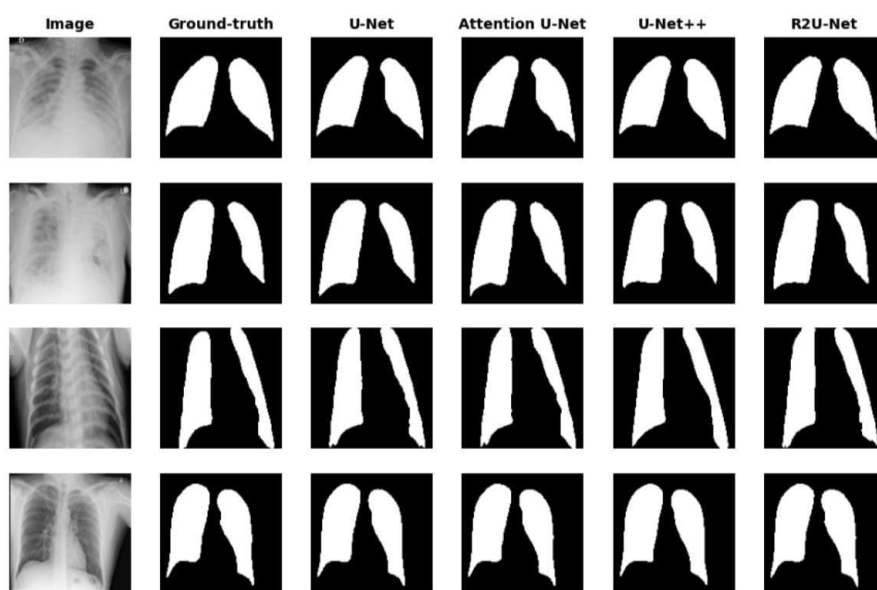
Mô hình	Tham số huấn luyện	Thời gian huấn luyện	Epoch
U-Net	2,158,417	27 phút	50
Attention U-Net	2,185,917	28 phút	50
U-Net++	2,478,853	27 phút	50
R2U-Net	6,373,169	51 phút	50

Bảng 5. Mô tả kết quả phân đoạn trên tập kiểm thử ảnh X-quang lồng ngực

Mô hình	Loss	DSC	IoU	Accuracy	Precision	Recall	F1-Score
U-Net	0.026	0.9801	0.9618	0.991	0.9819	0.9789	0.9801
Attention U-Net	0.0284	0.9785	0.9589	0.9903	0.983	0.9747	0.9785
U-Net++	0.0343	0.9725	0.9478	0.9876	0.9783	0.9674	0.9725
R2U-Net	0.0233	0.9825	0.9663	0.9921	0.9843	0.9814	0.9825

Các đám mờ trên ảnh X-quang là yếu tố gây nhiễu thông tin đối với việc phân đoạn vùng phổi. Đặc biệt đối với các bệnh nhân nặng, các đám mờ này nằm chồng lấp cấu trúc phổi gây khó khăn cho việc phân đoạn, nhưng hình ảnh lại có độ đồng nhất và tương phản cao hơn làm giảm độ phức tạp trong phân đoạn. Bảng 2 trình bày hiệu năng trên tập kiểm thử. R2U-Net đứng đầu với DSC đạt 0.98, IoU đạt 0.96 nhờ Recurrent Convolutional Layer và Residual Block học chi tiết sâu, giảm nhiễu và mất mát thông tin ngay cả khi có đám mờ. U-Net với IoU đạt 0.96, DSC đạt 0.98 kém hơn R2U do thiếu cơ chế học ngữ cảnh. Attention U-Net bị ảnh hưởng bởi đám mờ dày đặc, khi đó AG sẽ loại bỏ nhầm vùng phổi (Recall đạt 0.97). U-Net++ có hiệu năng thấp nhất (DSC đạt 0.97, IoU đạt 0.94) do khối tích chập dày đặc học không hiệu quả khi nhiễu che khuất cấu trúc vùng phổi

Attention U-Net sử dụng Cổng chú ý để giảm nhiễu nhưng với đám mờ dày đặc, mô hình có thể loại bỏ nhầm một số vùng phổi khi phân đoạn. Ngoài ra, yếu tố nhiễu khá lớn sẽ ảnh hưởng đến khả năng học thông tin của các lớp tích chập dày đặc, dẫn đến hiệu năng phân đoạn của U-Net++ không được cải thiện đáng kể. R2U-Net hoạt động tốt nhất trên tập dữ liệu này nhờ sử dụng kiến trúc khối tích chập hồi tiếp - phần dư, giúp mô hình có khả năng học sâu hơn các chi tiết của hình ảnh và giữ được thông tin quan trọng ngay cả khi đám mờ che khuất vùng phổi.



Hình 11: Một số kết quả phân đoạn vùng phổi của các mô hình

4. KẾT LUẬN

Nghiên cứu đã phân tích tổng quan kiến trúc mô hình U-Net cùng với các biến thể Attention U-Net, U-Net++ và R2U-Net, cung cấp cơ sở khoa học về đặc điểm và cấu trúc của từng mô hình. Hiệu năng của các mô hình được đánh giá thông qua các độ đo Loss, Dice Score, IoU, Accuracy, Precision, Recall và F1-Score trên hai tập dữ liệu vùng tổn thương da và X-quang lồng ngực. Kết quả cho thấy hiệu suất của mỗi mô hình phụ thuộc đáng kể vào đặc trưng của tập dữ liệu, bao gồm độ phức tạp cấu trúc và mức độ nhiễu. Việc đánh giá trên hai tập dữ liệu khác nhau đã làm nổi bật sự khác biệt này, giúp xác định mô hình phù hợp cho từng bài toán phân đoạn. Trong tương lai, các kiến trúc Transformer-base U-Net hoặc mô hình kết hợp nhiều biến thể sẽ được tích hợp để cải thiện độ chính xác phân đoạn trong điều kiện dữ liệu hạn chế. Nghiên cứu này tập trung vào phân tích định tính và so sánh trực quan hiệu năng giữa các mô hình. Các kiểm định thống kê như Wilcoxon hoặc Kruskal–Wallis không được áp dụng trong phạm vi hiện tại nhưng sẽ là hướng mở trong các nghiên cứu tiếp theo nhằm xác lập tính khác biệt có ý nghĩa thống kê.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), mã số VINIF.2024.ThS.36.

TÀI LIỆU THAM KHẢO

- [1]. D. Ciresan, A. Giusti, L. M. Gambardella and J. Schmidhuber (2012). Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images, *Advances in Neural Information Processing Systems*, vol. 25.
- [2]. E. Shelhamer, J. Long and T. Darrel (2016). Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 640-651.
- [3]. M. Liang, X. Hu (2015). Recurrent Convolutional Neural Network for Object Recognition, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, pp. 3367-3375.
- [4]. K. He, X. Zhang, S. Ren and J. Sun (2016). Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 770-778.
- [5]. O. Ronneberger, P. Fischer and T. Brox (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, Munich, Germany, vol. 9351, pp. 234-241.
- [6]. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al (2018). Attention U-Net: Learning Where to Look for the Pancreas, *MIDL 2018 - Medical Imaging with Deep Learning*, Amsterdam, Netherlands.
- [7]. X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng (2018). H-DenseUNet: Hybrid Densely Connected U-Net for Liver and Tumor Segmentation from CT Volumes, *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2663-2674.
- [8]. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang (2019). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, *IEEE Transactions on Medical Imaging*, vol. 39, pp. 1856-1867.
- [9]. M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari (2019). Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation, *Journal of Medical Imaging*, vol. 6, pp. 014006.
- [10]. N. Siddique, S. Paheding, C. P. Elkin and V. Devabhaktuni (2021). U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications, *IEEE Access*, vol. 9, pp. 82031-82057.
- [11]. R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval and A. Bozorgpour (2024). Medical Image Segmentation Review: The Success of U-Net, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 10076-10095.
- [12]. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia (2017). Attention Is All You Need, *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 6000-6010.

- [13]. A. Galassi, M. Lippi, P. Torroni (2021). Attention in Natural Language Processing, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 4291-4308.
- [14]. M. H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu (2022). Attention mechanisms in computer vision: A survey, *Computational Visual Media*, vol. 8, pp. 331-368.
- [15]. W. A. Albukhanajer, Y. Jin, J. A. Briffa (2015). Trade-off Between Computational Complexity and Accuracy in Evolutionary Image Feature Extraction, *2015 IEEE Congress on Evolutionary Computation (CEC)*, Sendai, Japan, pp. 2412-2419.

PERFORMANCE EVALUATION OF U-NET FAMILY MODELS IN MEDICAL IMAGE SEGMENTATION

Vo Van Thanh^{1*}, Le Sy Toan²

¹ Faculty of Information Technology, University of Science, Hue University

² K45F, Faculty of Information Technology, University of Science, Hue University

*Email: vvthanh@husc.edu.vn

ABSTRACT

Medical image segmentation is a critical application in deep learning, playing a pivotal role in supporting diagnosis, treatment, and disease analysis in the healthcare domain. In this paper, we compare the performance of several U-Net family models, including U-Net, Attention U-Net, U-Net++, and R2U-Net on two medical imaging datasets: dermoscopic lesion images and chest X-ray images. The evaluation is based on standard metrics such as DSC and IoU. Results show that R2U-Net consistently delivers the highest segmentation accuracy (DSC > 0.91 on the dermoscopic lesion dataset and DSC > 0.98 on the chest X-ray dataset). At the same time, the relative performance of the other models varies depending on image characteristics. This study provides a foundation for selecting the most suitable U-Net variant to enhance the accuracy of medical image diagnosis. All source code and dataset details are publicly available at <https://github.com/LeSyToan2003/Medical-Segmentation.git>.

Keywords: Medical Image Segmentation, Deep Learning, U-Net, U-Net Variants, Convolutional Neural Network (CNN).



Võ Văn Thành sinh năm 2000 tại thành phố Huế. Ông tốt nghiệp Cử nhân ngành Công nghệ thông tin năm 2022 tại trường Đại học Khoa học, Đại học Huế. Ông hiện đang là trợ giảng Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Học máy, học sâu, thị giác máy tính.