Medical Image Segmentation: A project

Lê Phạm Hoàng Trung
21120157
Khoa Công nghệ Thông tin
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
Thành phố Hồ Chí Minh, Việt Nam
21120157@student.hcmus.edu.vn
0985879454

Lê Văn Tấn
21120554
Khoa Công nghệ Thông tin
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
Thành phố Hồ Chí Minh, Việt Nam
21120554@student.hcmus.edu.vn
0944745303

Nguyễn Quang Vinh
21120604
Khoa Công nghệ Thông tin
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
Thành phố Hồ Chí Minh, Việt Nam
21120604@student.hcmus.edu.vn
0949245895

***Abstract*—[Tóm tắt] Trong lĩnh vực y học hiện đại, việc phân đoạn hình ảnh y khoa đóng một vai trò quan trọng trong việc chẩn đoán và điều trị bệnh. Các nghiên cứu gần đây đã đưa ra nhiều mô hình phân đoạn hình ảnh y khoa, từ các phương pháp truyền thống đến học sâu. Bài viết này khảo sát và đề xuất một mô hình sử dụng kiến trúc U-Net, sau đó kết hợp các kỹ thuật trong xử lý hình ảnh và học sâu, nhằm cải thiện độ chính xác trong việc phân đoạn hình ảnh y khoa.**

***Index Terms*—medical image segmentation, deep learning, U-Net**

I. Introduction

Trong những năm gần đây, học sâu đã trở thành một công cụ không thể thiếu trong lĩnh vực phân đoạn hình ảnh y tế, mang lại những cải tiến đáng kể về độ chính xác và hiệu quả chẩn đoán. Các mạng lưới nơ-ron tích chập (CNN) đã chứng minh khả năng đại diện cho các đặc trưng hình ảnh một cách phân cấp, từ đó trở thành chủ đề nghiên cứu sôi nổi trong xử lý ảnh và thị giác máy tính. Sự không nhạy cảm với nhiễu ảnh, mờ, độ tương phản, v.v., của CNNs cung cấp kết quả phân đoạn xuất sắc cho hình ảnh y tế.

Phân đoạn hình ảnh y tế nhằm làm rõ hơn sự thay đổi của cấu trúc giải phẫu hoặc bệnh lý trong hình ảnh; nó thường đóng vai trò quan trọng trong chẩn đoán hỗ trợ máy tính và y học thông minh. Các nhiệm vụ phân đoạn hình ảnh y tế phổ biến bao gồm phân đoạn gan và khối u gan, não và khối u não, đĩa quang học, tế bào, phổi, nốt phổi, hình ảnh tim, v.v. Với sự phát triển và phổ biến của thiết bị hình ảnh y tế, X-quang, CT, MRI và siêu âm đã trở thành bốn phương tiện hỗ trợ hình ảnh quan trọng giúp các bác sĩ chẩn đoán bệnh, đánh giá tiên lượng và lập kế hoạch phẫu thuật tại các cơ sở y tế.

Căn cứ vào số lượng dữ liệu được gán nhãn, các phương pháp cho phân đoạn ảnh y tế được chia thành ba nhóm chính là học có giám sát, học giám sát yếu và học không giám sát. Ưu điểm của việc học có giám sát là có thể huấn luyện đầy đủ trên bộ dữ liệu có nhãn, nhưng có một điều là rất khó để có được một lượng lớn dữ liệu được gán nhãn cho hình ảnh y tế. Việc gán nhãn cho hình ảnh y tế đòi hỏi độ chính xác và mức độ hiểu biết cao, điều này là không hề đơn giản trong thực tế. Ngược lại, học không giám sát không yêu cầu dữ liệu có nhãn, nhưng độ khó của việc học sẽ được tăng lên, do độ phức tạp của hình ảnh y tế. Học giám sát yếu nằm giữa học có giám sát và học không giám sát vì nó chỉ yêu cầu một phần nhỏ dữ liệu được gán nhãn.

Báo cáo sẽ trình bày một số mô hình tiêu biểu để cho thấy sự phát triển trong phân đoạn ảnh y tế. Trong đó tập trung vào học có giám sát và học giám sát yếu là chủ yếu bởi vì các mô hình không giám sát hiện tại cho kết quả không được cạnh tranh so với hai phương pháp còn lại và hiện tại không được phổ biến rộng rãi. Sau khi đã tiến hành khảo sát, báo cáo sẽ đề xuất một mô hình học giám sát yếu cho phân đoạn ảnh y khoa và sau đó thực hiện đánh giá cũng như cải tiến trên mô hình này.

II. Related work

A. Supervised learning

1) Backbone: Phân đoạn ngữ nghĩa hình ảnh đó là nhằm mục đích phân loại từng pixel của ảnh. Với mục tiêu này, kiến trúc encoder-decoder được đề xuất với một số kiến trúc phổ biến như Fully Convolutional Network (FCN) [1], U-Net [2], v.v. Trong các kiến trúc này bộ encoder thường được sử dụng để trích xuất các đặc trưng hình ảnh, trong khi decoder được sử dụng để khôi phục các đặc trưng để khôi phục lại ảnh có kích thước ban đầu và xuất ra kết quả phân đoạn. Một trong những kiến trúc có độ ảnh hưởng lớn đó là U-Net.

U-Net: Unet được phát triển dựa trên cơ sở mạng tích chập đầy đủ (FCN), nhằm giải quyết vấn đề hiệu suất do phải lấy tích chập ảnh với độ phân giải không đổi xuyên suốt các lớp mạng. Unet gồm có ba phần chính là bộ mã hoá (đường rút gọn), bộ giải mã (đường mở rộng) và liên kết trực tiếp.

Bộ mã hoá có kiến trúc tương tự các mạng nơ-ron tích chập gồm các phép tích chập, hàm kích hoạt và phép tổng hợp bằng giá trị lớn nhất cho bước lấy mẫu xuống. Tại mỗi bước lấy mẫu xuống, tổng số kênh đặc trưng tăng lên đồng thời độ phân giải giảm xuống.

Bộ giải mã là điểm mới của Unet so với FCN, gồm có bước lấy mẫu lên bằng phép tích chập, phép nối bản đồ đặc trưng từ bộ mã hoá sang bản đồ đặc trưng hiện tại (thông qua liên kết trực tiếp) và phép tích chập kết hợp hàm kích hoạt. Tại mỗi bước lấy mẫu lên, tổng số kênh đặc trưng sẽ giảm xuống, song độ phân giải tăng lên. Ở bước cuối, người ta sử dụng phép tích chập 1x1 để ánh xạ véc-tơ đặc trưng của từng điểm ảnh vào lớp phân loại phù hợp.

Ưu điểm của Unet là huấn luyện nhanh, cần ít dữ liệu gán nhãn và dễ áp dụng cho nhiều tác vụ. Tuy nhiên, Unet gặp phải vấn đề về độ sâu của mạng và khả năng trích xuất thông tin tầm xa.

2) Connection: Kết nối dày đặc thường được sử dụng để xây dựng một loại mạng nơ-ron tích chập đặc biệt. Đối với các mạng kết nối dày đặc, đầu vào của mỗi lớp đến từ đầu ra của tất cả các lớp trước đó trong quá trình truyền chuyển tiếp. Lấy cảm hứng từ kết nối dày đặc, người ta đã cải tiến U-Net bằng cách thay các khối của U-Net bằng các kết nối dày đặc.

UNet++: Trong UNet++ [3], người ta đã thiết kế lại đường kết nối giữa khối "encoder" và "decoder". Trong U-Net, các đặc trưng của bộ "encoder" được nhận trực tiếp trong bộ "decoder"; tuy nhiên, trong UNet++, chúng phải qua một khối tích chập lồng nhau dày đặc. Những đường dẫn giữa hai khối "encoder" và "decoder" này được gọi là "skip connection" vì chúng cho phép thông tin bỏ qua các lớp nhất định trong mạng, hỗ trợ luồng thông tin giữa các thành phần bộ "encoder" và bộ "decoder".

Mục tiêu chính đằng sau việc kết hợp các "skip connection" lồng nhau dày đặc trong UNet++ là để nâng cao độ chính xác phân đoạn hình ảnh y tế. Bằng cách dần dần làm phong phú các bản đồ đặc trưng ("feature maps") có độ phân giải cao từ mạng "encoder" trước khi hợp nhất chúng với các "feature maps" giàu ngữ nghĩa từ mạng "decoder", mô hình nhằm mục đích nắm bắt các chi tiết của các đối tượng tiền cảnh một cách hiệu quả hơn.

Một cách chính thức, chúng ta sẽ xây dựng hàm cho "skip connection" như sau: cho $x^{i,j}$ là biểu diễn đầu ra của một nút cụ thể trong mạng với i là chỉ mục cho lớp "down-sampling" trong bộ "encoder" (đường chéo xuống) và j là chỉ mục cho lớp tích chập dày đặc dọc theo đường "skip connection" (đường ngang). Cộng dồn ("stack") các "feature maps" dựa vào các đầu vào nó nhận được được biểu thị bởi $x^{i,j}$ qua công thức

$$x^{i,j} = \begin{cases} H(x^{i-1,j}), & j = 0 \\ H\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, U(x^{i+1,j-1})\right]\right), & j > 0 \end{cases} \quad (1)$$

Đối với các nút ở mức $j = 0$, "feature maps" được tính bằng hàm $H(x^{i-1,j})$, trong đó $H(.)$ biểu thị phép toán tích chập theo sau là hàm kích hoạt. Đối với các nút ở mức $j > 0$, "feature maps" được tính toán khác nhau. Nó liên quan đến việc ghép các "feature maps" từ lớp trước (được lập chỉ mục bằng k) và áp dụng thao tác "up-sampling" được ký hiệu là $U(.)$ để kết hợp thông tin từ mạng con của "decoder". Và [] là lớp kết nối.

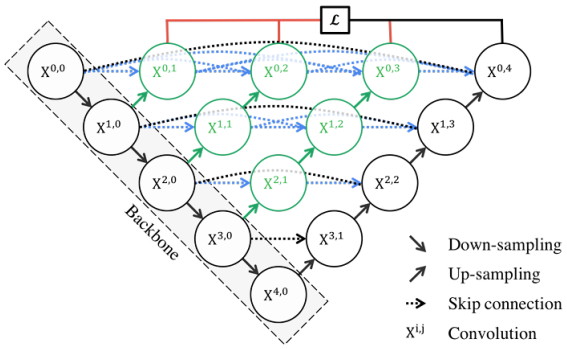

Fig. 1. Kiến trúc của UNet++

Về cơ bản, các nút ở mức $j = 0$ chỉ nhận được một đầu vào từ lớp trước của bộ mã hóa; các nút ở cấp độ $j = 1$ nhận hai đầu vào, cả hai đều từ mạng con của "encoder" nhưng ở hai cấp độ liên tiếp; và các nút ở cấp $j > 1$ nhận đầu vào $j + 1$, trong đó j đầu vào là đầu ra của j nút trước đó trong cùng một đường "skip connection" và đầu vào cuối cùng là đầu ra của "up-sampling" từ đường "skip connection" bên dưới (đường chéo lên).

Mặc dù kết nối dày đặc rất hữu ích để thu được các đặc trưng hình ảnh phong phú hơn nhưng nó thường làm giảm độ mạnh của việc biểu diễn đặc trưng ở một mức độ nhất định và làm tăng số lượng tham số.

3) Transformer Mechanism: Với sự phát triển đột phá của Transformer trong lĩnh vực NLP, người ta mong muốn áp dụng kiến trúc đó lên lĩnh vực xử lý ảnh. Do đó một nhóm tác giả đã đề xuất một mô hình gọi là TransUnet [4].

TransUnet: TransUnet là một mô hình kết hợp cả hai công nghệ là Transformers và U-Net để phân đoạn hình ảnh y khoa, sự kết hợp của hai công nghệ này giúp cải thiện khả năng phân đoạn hình ảnh y khoa. Transformers được sử dụng như các bộ encoders để xử lý bối cảnh toàn cục của hình ảnh, trong khi U-Net được sử dụng để định vị chính xác các vùng quan trọng trong hình ảnh. Mô hình đạt được hiệu suất ưu việt bằng cách mã hóa các điểm ảnh của hình ảnh đã được mã hóa từ một bản đồ đặc trưng CNN và kết hợp chúng với các bản đồ đặc trưng CNN có độ phân giải cao. Bên cạnh đó, TransUNet cũng vượt qua các mô hình trước đó khi so sánh với chúng trong cùng lĩnh vực và đạt được kết quả tốt nhất trong các biến thể khác nhau của các mô hình dựa trên Transformer, thiết lập một new state-of-the-art trong phân đoạn hình ảnh y khoa.

B. Weakly supervised learning

1) Data Augmentation: Trong trường hợp không có các bộ dữ liệu được dán nhãn lớn, việc tăng cường dữ liệu là một giải pháp hiệu quả cho vấn đề này. Tuy nhiên, các phương pháp tăng cường dữ liệu thông thường tạo ra hình ảnh có độ tương quan cao với hình ảnh gốc. So với các phương pháp tăng cường dữ liệu phổ biến, GAN [5] do Goodfellow đề xuất

hiện là một chiến lược phổ biến để tăng cường dữ liệu vì GAN khắc phục được vấn đề phụ thuộc vào dữ liệu gốc.

Phương pháp truyền thống: Các phương pháp tăng cường dữ liệu thông thường bao gồm cải thiện chất lượng hình ảnh như khử nhiễu, thay đổi cường độ hình ảnh như độ sáng, độ bão hòa và độ tương phản cũng như thay đổi bố cục hình ảnh như xoay, biến dạng và chia tỷ lệ, v.v. Ngoài ra còn có một số phương pháp cải tiến hơn như Gaussian blur, sử dụng ngẫu nhiên hàm tăng cường độ sáng trong hình ảnh MR 3D để làm đa dạng dữ liệu huấn luyện, ...

Conditional Generative Adversarial Nets (cGAN) [6]: Trình tạo sinh GAN gốc, ký hiệu là G có thể học cách phân phối dữ liệu, nhưng các hình ảnh được tạo ra là ngẫu nhiên, điều đó có nghĩa là quá trình G là trạng thái không được hướng dẫn, cGAN thêm một điều kiện vào GAN ban đầu để hướng dẫn quá trình G.

2) Transfer Learning: Bằng cách sử dụng các tham số đã được huấn luyện của mô hình để khởi tạo một mô hình mới, việc học chuyển giao có thể đạt được việc huấn luyện mô hình nhanh chóng cho dữ liệu có nhãn hạn chế. Một cách tiếp cận là tinh chỉnh mô hình được đào tạo trước trên ImageNet cho nhiệm vụ phân tích trên hình ảnh y tế mục tiêu.

Pre-trained model: Học chuyển giao thường được sử dụng để giải quyết vấn đề hạn chế dữ liệu được gắn nhãn trong phân tích hình ảnh y tế và một số nhà nghiên cứu nhận thấy rằng việc sử dụng các mạng được đào tạo trước trên các hình ảnh tự nhiên như ImageNet làm bộ "encoder" trong mạng có hình dáng tương tự U-Net và sau đó thực hiện "fine-tuning" trên dữ liệu y tế, điều này có thể cải thiện hiệu quả phân đoạn của hình ảnh y tế.

III. Proposed Method

Mô hình U-Net đã chứng minh khả năng học hiệu quả từ một tập huấn luyện tương đối nhỏ. Điều này đặc biệt quan trọng trong ngữ cảnh phân đoạn hình ảnh y khoa, nơi mà việc chuẩn bị dữ liệu thủ công có thể tốn kém và mất thời gian. U-Net thường được đào tạo từ đầu, với các trọng số được khởi tạo một cách ngẫu nhiên. Điều này giúp tránh hiện tượng quá khớp khi huấn luyện trên tập dữ liệu lớn, có thể lên đến hàng triệu hình ảnh. Một phương pháp phổ biến khác là sử dụng các mạng đã được đào tạo trên tập dữ liệu Imagenet như một điểm khởi đầu cho việc huấn luyện. Cách tiếp cận này cho phép quá trình học tập được tinh chỉnh cho một số lớp mạng chưa được đào tạo trước, thường là lớp cuối cùng, để tận dụng các đặc trưng cụ thể của dữ liệu đang được xử lý.

Dựa vào những khảo sát phía trên, ứng dụng của mạng học sâu đã được đào tạo vào ảnh y khoa [7], và với tiêu chí phù hợp với phạm vi đồ án môn học, nhóm chúng em đề xuất mô hình cải tiến của kiến trúc U-Net. Chúng em sử dụng mạng có cấu trúc sử dụng bộ "encoder" được đào tạo trước. TernausNet [8] là một kiến trúc giống U-Net sử dụng các mạng VGG11 [9] được đào tạo trước tương đối đơn giản làm bộ "encoder" với trọng số được huấn luyện sẵn trên bộ dữ liệu lớn ImageNet. VGG11 bao gồm 7 lớp tích chập, mỗi lớp theo sau là hàm kích hoạt ReLU, cùng với đó 5 max pooling 2 x 2. Tất cả các lớp chập đều có kernel 3 × 3 và số lượng channel được cho bởi

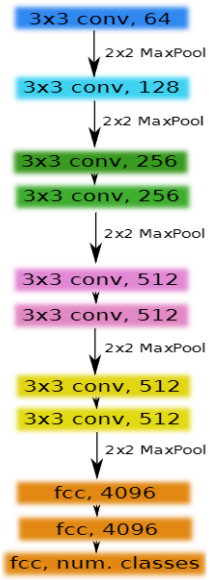

Fig. 2. Kiến trúc mạng VGG11

Bộ dữ liệu	Số lượng ảnh	Kích thước	Nhà cung cấp
PH2	200	768x560	Dự án ADDI
Kvasir	1000	720x676 → 1920x1072	Simula

TABLE I
Những bộ dữ liệu phân đoạn ảnh y tế được dùng để thử nghiệm

Hình 2. Để xây dựng một bộ "encoder", ta loại bỏ các lớp kết nối đầy đủ và thay thế chúng bằng một lớp tích chập duy nhất với 512 channel, xem như phần trung tâm của mạng, tách bộ "encoder" khỏi bộ "decoder". Để xây dựng bộ "decoder", sử dụng các lớp tích chập chuyển vị cho phép tăng gấp đôi kích thước của "feature map" trong khi đó giảm số lượng channel một nửa. Đầu ra của một tích chập chuyển vị sau đó được nối với đầu ra của phần tương ứng của bộ "encoder". "Feature map" kết quả được xử lý bằng thao tác tích chập để giữ cho số lượng channel giống như đầu ra của bộ "encoder" tương ứng. Thủ tục này được lặp lại 5 lần để ghép cặp với 5 lớp max pooling ở "encoder", như được hiển thị trong Hình 3. Hiện tại, vì ta có 5 lớp max pooling, mỗi lần giảm kích thước đi hai lần, nên chỉ có hình ảnh có kích thước chia hết cho 32 (2^5) mới có thể được sử dụng làm đầu vào cho mạng hiện tại.

IV. Experiments

A. Datasets

Bảng I tóm tắt thông tin về hai tập dữ liệu được sử dụng để đánh giá mô hình. Mô tả chi tiết từng tập dữ liệu sẽ được trình bày sau đây.

PH2: Hình ảnh soi da được thu tập tại Khoa Da liễu của Bện viện Pedro Hispano (Matosinhos, Bồ Đào Nha) trong cùng điều kiện thông qua hệ thống Máy phân tích nốt ruồi Tuebinger với độ phóng đại 20x. Tất cả được thu thập dưới dạng ảnh màu RGB 8-bit với độ phân giải 768x560 pixel.

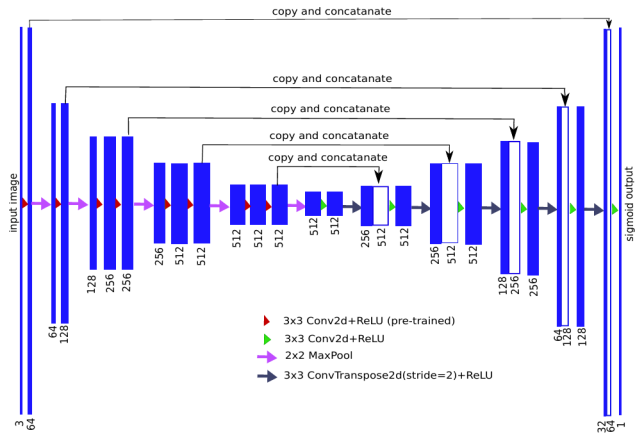

Fig. 3. Ternausnet với VGG11 cho phân đoạn nhị phân. Mỗi khối hình chữ nhật màu xanh biểu diễn một "feature map" nhiều channel đi qua một loạt các biến đổi. Chiều cao của thanh biểu thị kích thước tương đối của "feature map" (theo pixel), trong khi chiều rộng của chúng tỷ lệ với số lượng kênh (số lượng được ghi rõ ràng dưới dạng chỉ số cho thanh tương ứng). Số lượng kênh tăng theo từng giai đoạn ở phần bên trái trong khi giảm theo từng giai đoạn ở phần "decoder" bên phải. Các mũi tên ở phía trên cho thấy việc chuyển giao thông tin từ mỗi lớp "encoder" và nối nó với lớp "decoder" tương ứng

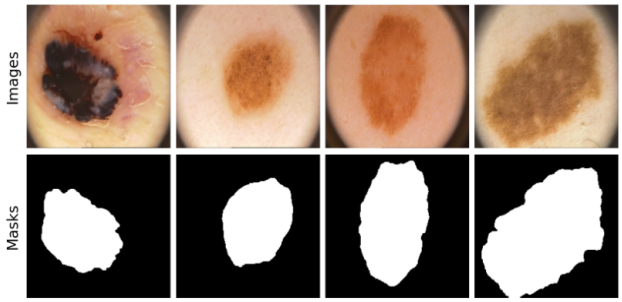

Fig. 4. Một vài ảnh và mặt nạ của bộ dữ liệu PH2

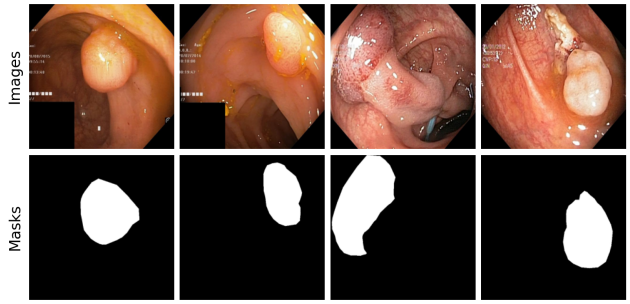

Fig. 5. Một vài ảnh và mặt nạ của bộ dữ liệu Kvasir

Bộ dữ liệu	Tổng số lượng ảnh	Train	Val	Test
PH2	200	100	50	50
Kvasir	1000	600	200	200

TABLE II
SỐ LƯỢNG ẢNH TRONG MỖI TẬP TRÊN HAI BỘ DỮ LIỆU

Tập dữ liệu hình ảnh PH2 chứa tổng cộng 200 hình ảnh soi da của các tổn thương tế bào hắc tố, bao gồm 80 nốt ruồi thông thường, 80 nốt ruồi không điển hình và 40 khối u ác tính. Tập dữ liệu gồm các hình ảnh được đánh nhãn, cụ thể là phân đoạn y khoa của tổn thương, chẩn đoán lâm sàng và mô học cũng như đánh giá một số tiêu chí soi da (màu sắc; mạng lưới sắc tố; chấm/ hạt; vệt; vùng hồi quy; tấm màn trắng xanh).
Hình 4 minh hoạ vài mẫu trong tập dữ liệu PH2.

Kvasir: Là một tập dữ liệu lớn và đa dạng được sử dụng cho các bài toán phân đoạn hình ảnh trong lĩnh vực y tế, đặc biệt là phân đoạn polyp đại tràng. Bộ dữ liệu này được thu thập và chú thích bởi các chuyên gia y tế, cung cấp một nguồn tài liệu quý giá cho việc phát triển và đánh giá các thuật toán phân đoạn polyp tự động. Gồm hơn 1000 hình ảnh nội soi đại tràng có độ phân giải cao, được chụp từ nhiều bệnh nhân khác nhau với nhiều tình trạng đại tràng khác nhau. Mỗi hình ảnh được chú thích thủ công bằng mặt nạ phân đoạn, xác định vị trí và ranh giới của các polyp đại tràng với độ chính xác cao. Các chú thích được thực hiện bởi các chuyên gia y tế có kinh nghiệm, đảm bảo độ tin cậy và tính nhất quán cho bộ dữ liệu.

Cả hai bộ dữ liệu đều được chia ra thành 3 phần train, val, test để phục vụ cho quá trình huấn luyện và đánh giá. Bảng II thể hiện số lượng mẫu cho mỗi tập.

B. Loss function

Do cả hai tập dữ liệu PH2 và Kvasir đều cung cấp nhãn phân đoạn dưới dạng nhị phân, hàm mất mát Binary Cross Entropy Loss được sử dụng trong quá trình huấn luyện mô hình. Hàm này đánh giá độ chính xác của các mặt nạ dự đoán so với mặt nạ thực tế. Công thức của hàm mất mát Binary Cross Entropy Loss được biểu diễn như sau:

$$\mathcal{L}_{BCE}(y, \hat{y}) = -\sum_{i} [y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \tag{2}$$

trong đó y_i mang giá trị nhị phân, là mặt nạ thực tế của pixel i và $\hat{y}_i$ là đầu ra của mô hình ứng với pixel i, σ tượng trưng cho hàm Sigmoid

$$\sigma(x) = \frac{e^x}{1 + e^x} \tag{3}$$

C. Metrics

Hai loại độ đo được sử dụng đó là IOU và Dice coefficient. Cụ thể hơn, IOU (Intersection Over Union) đo lường tỷ lệ giữa diện tích giao nhau và diện tích hợp của các hộp bao (bounding box) hoặc mặt nạ (mask) dự đoán và thực tế. IOU

có giá trị nằm trong khoảng từ 0 đến 1, trong đó 1 là khớp hoàn toàn và 0 là không có giao nhau.

$$\text{IOU} = \frac{\text{groundtruth} \cap \text{prediction}}{\text{groundtruth} \cup \text{prediction}} \tag{4}$$

Một độ đo phổ biến khác trong các bài toán phân đoạn ảnh đó là Dice coefficient, còn được gọi là Dice similarity coefficient, được tính toán theo công thức sau:

$$\text{Dice} = \frac{2 \times (\text{groundtruth} \cap \text{predict})}{\text{groundtruth} + \text{predict}} \tag{5}$$

Dice coefficient nhấn mạnh sự giao nhau hơn, do đó có khả năng phân biệt rõ ràng hơn giữa các trường hợp có sự trùng lặp cao và thấp, ngay cả khi kích thước của các mask khác nhau. Giả sử chúng ta có một hình ảnh phân đoạn não bộ, với vùng thực tế là khối u não và vùng dự đoán là kết quả phân đoạn từ mô hình. Dice Coefficient sẽ cao hơn nếu mô hình dự đoán được nhiều pixel bên trong khối u, ngay cả khi có một số pixel bị bỏ sót. IOU sẽ cao hơn nếu mô hình dự đoán bao phủ toàn bộ khối u, nhưng có thể bao gồm cả một số pixel không thuộc khối u.

D. Data Augmentation

Tăng cường dữ liệu là một phương pháp phổ biến nhằm nâng cao độ chính xác của mô hình thông qua việc tăng số lượng mẫu trong tập dữ liệu huấn luyện.

Trong đồ án này, nhóm đã thử nghiệm một số phương pháp tăng cường dữ liệu như lật ảnh theo chiều dọc, lật ảnh theo chiều ngang, xoay ảnh 90 độ, ... Tuy nhiên, kết quả độ chính xác của mô hình không có sự cải thiện.

E. Training

Để huấn luyện mô hình phân đoạn, chúng ta sử dụng thuật toán tối ưu hóa AdamW với tốc độ học tập ban đầu (learning rate) là 4e-5 và độ suy hao trọng số (weight decay) là 1e-4. Quá trình huấn luyện được thực hiện trong 50 epoch, sử dụng kiến trúc mạng nơ-ron TernausNet11 được mô tả chi tiết trong phần III. Hàm mất mát được sử dụng là BCELoss đã được định nghĩa phía trên.

Tiền xử lý dữ liệu đầu vào: Ảnh đầu vào được thay đổi kích thước thành (224x224) để phù hợp với kích thước đầu vào của mạng TernausNet11 và các giá trị được chuẩn hóa về đoạn $[0, 1]$. Đầu ra của mạng là một ma trận có kích thước tương tự như ảnh đầu vào, nhưng chỉ có một kênh duy nhất, thể hiện giá trị tại mỗi pixel, giá trị này sau đó sẽ được đưa vào hàm sigmoid để tính xác xuất pixel có thuộc mặt nạ hay không. Tương ứng, các mặt nạ groundtruth cũng được tiền xử lý để chỉ tồn tại các giá trị nhị phân 0, 1.

Tính toán xác suất pixel: Giá trị đầu ra của mạng được đưa vào hàm sigmoid để chuyển đổi thành giá trị xác suất nằm trong khoảng từ 0 đến 1. Hàm sigmoid được tích hợp sẵn trong hàm nn.BCEWithLogitsLoss() của PyTorch, do đó, chỉ cần đưa giá trị đầu ra của mô hình vào hàm loss này.

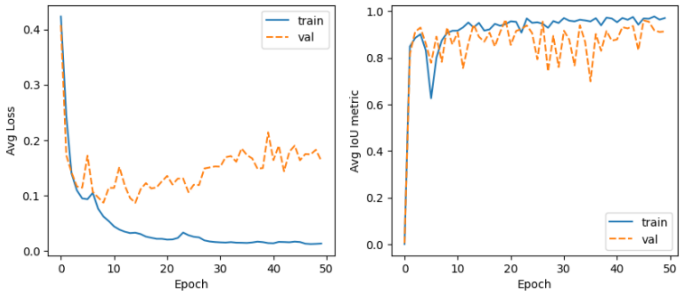

Fig. 6. Quá trình huấn luyện của TernauNet11 trên PH2

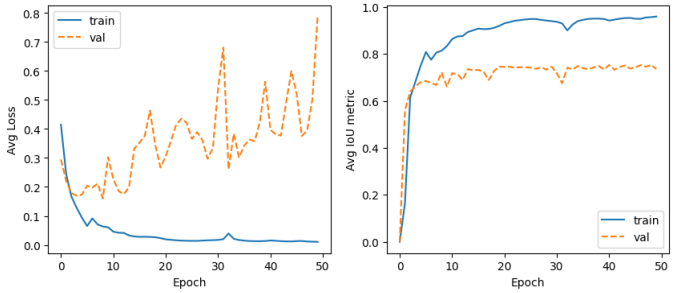

Fig. 7. Quá trình huấn luyện của TernauNet11 trên Kvasir

Mục tiêu tối ưu hóa: Mục tiêu huấn luyện là tối thiểu hóa hàm mất mát BCELoss, sao cho giá trị dự đoán của mô hình (xác suất pixel) càng gần với giá trị của mặt nạ groundtruth (mặt nạ thực tế) càng tốt. Quá trình tối ưu hóa được thực hiện bằng cách cập nhật trọng số của mạng nơ-ron theo hướng giảm dần giá trị hàm mất mát.

Chuyển đổi xác suất dự đoán thành mặt nạ nhị phân: Mặc dù xác suất dự đoán được sử dụng trong hàm mất mát để tối ưu hóa mô hình, để đánh giá độ chính xác của mô hình một cách trực quan và định lượng, chúng ta cần chuyển đổi xác suất dự đoán thành mặt nạ nhị phân. Những pixel có giá trị lớn hơn hoặc bằng 0.5 được gán giá trị 1, tương ứng với màu trắng trong ảnh, thể hiện rằng mô hình dự đoán có sự hiện diện của mặt nạ tại vị trí đó. Ngược lại, những pixel có giá trị nhỏ hơn 0.5 được gán giá trị 0, tương ứng với màu đen trong ảnh, thể hiện rằng mô hình dự đoán không có mặt nạ tại vị trí đó. Chúng ta sẽ dùng mặt nạ nhị phân này để tính các độ đo mà chúng ta đã định nghĩa phía trên.

Quá trình huấn luyện được thể hiện như hình 6 và hình 7.

F. Results

Sau khi hoàn tất quá trình huấn luyện, mô hình được đánh giá trên tập thử nghiệm đã được định nghĩa trước đó. Việc đánh giá này nhằm mục đích kiểm tra hiệu suất của mô hình trên dữ liệu mới, chưa từng được sử dụng trong quá trình huấn luyện. Sau đó là so sánh hiệu suất của mô hình với các công trình liên quan trong lĩnh vực phân đoạn ảnh y tế.

1) PH2: Việc so sánh được thực hiện dựa trên các benchmark được cung cấp bởi trang web Paper With Code https://paperswithcode.com/dataset/ph2. Bảng III thể hiện kết quả

Mô hình	mIOU	meanDice
SegNet	0.9361	
MFSNet	0.9140	0.9540
DermoSegDiff-B		0.9467
TernausNet11	0.8660	0.9250

TABLE III
KẾT QUẢ TRÊN BỘ DỮ LIỆU PH2 VÀ SO SÁNH VỚI CÁC PHƯƠNG PHÁP KHÁC

Mô hình	mIOU	meanDice
U-Net		0.818
U-Net++		0.821
ResUNet		0.7877
ResUNet++		0.8133
TransNetR	0.8016	0.8706
TransResU-Net	0.8214	0.8884
FCB-SwinV2 Transformer	0.8973	0.9420
TernausNet11	0.781	0.857

TABLE IV
KẾT QUẢ TRÊN BỘ DỮ LIỆU KVASIR VÀ SO SÁNH VỚI CÁC PHƯƠNG PHÁP KHÁC

của mô hình được đề xuất và so sánh nó với các công trình nghiên cứu liên quan. Kết quả cho thấy mô hình Ternaunet11 có vẻ chưa phù hợp với tập dữ liệu PH2 vốn có số lượng ảnh hạn chế.

2) Kvasir: Việc so sánh được thực hiện trên benchmark Medical Image Segmentation on Kvasir-SEG được cung cấp bởi Paper With Code https://paperswithcode.com/sota/medical-image-segmentation-on-kvasir-seg. Bảng IV thể hiện kết quả của mô hình được đề xuất và so sánh nó với các công trình nghiên cứu liên quan, ở đây chỉ nêu ra một vài nghiên cứu tiêu biểu. Kết quả cho thấy mô hình cũng đã vượt qua một vài phương pháp truyền thống và lép vế so với các phương pháp hiện đại hiện nay, nhìn chung đây cũng là kết quả đáng kể cho một mô hình mạng có phần đơn giản.

V. ADVANCE

Kết quả cho thấy việc sử dụng bộ mã hóa VGG11, vốn có kiến trúc tương đối nông và khả năng trích xuất đặc trưng hạn chế. Hiện nay có nhiều phiên bản cải tiến hơn của các mô hình pretrained VGG, nhóm đề xuất cải tiến đó là sử dụng một mạng có kiến trúc sâu hơn đó là VGG16 làm bộ encoder thay cho VGG11. Như tên gọi, VGG16 có tổng 16 lớp tích chập và mạng kết nối đầy đủ. Hình 8 cho ta cái nhìn tổng quan về kiến trúc 2 mạng VGG11 và VGG16, thứ mà sẽ dùng để làm bộ encoder cho mô hình được đề xuất. Dễ quan sát nhất đó là với mỗi cụm mạng VGG16 thêm một lớp tích chập so với mạng VGG11. VGG16 có kiến trúc sâu hơn và

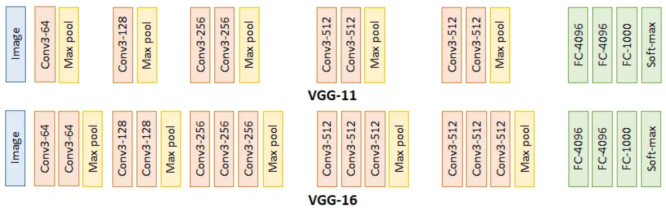

Fig. 8. So sánh giữa VGG11 và VGG16

Mô hình	Số lượng tham số
TernausNet11	32.15M
TernausNet16	44.02M

TABLE V
SỐ LƯỢNG THAM SỐ CỦA HAI MÔ HÌNH TERNAUSNET11 VÀ TERNAUSNET16

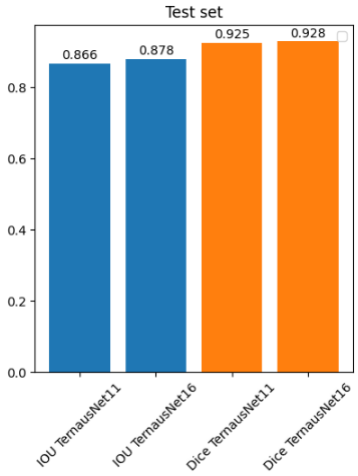

Fig. 9. So sánh kết quả của TernausNet11 và TernausNet16 trên PH2

khả năng trích xuất đặc trưng tốt hơn, có số lượng tham số nhiều hơn được thể hiện ở bảng V, hi vọng sẽ đạt hiệu suất phân đoạn cao hơn trên các tập dữ liệu benchmark.

Để đánh giá hiệu suất cải tiến, chúng em cũng đã thực nghiệm trên hai tập dữ liệu phía trên và cho kết quả như hình 9 và hình 10. Kết quả thực nghiệm cho thấy mô hình đã được cải tiến thể hiện sự tăng về độ chính xác trên cả hai tập dữ liệu khảo sát, đạt hiệu suất cao hơn trên cả hai chỉ số meanIOU và Dice mean. So với các công trình nghiên cứu liên quan, mô hình này tuy chưa vượt qua hiệu suất của các mô hình trên tập dữ liệu PH2, nhưng đã thể hiện khả năng cạnh tranh cao trên tập dữ liệu Kavir, sánh ngang với mô hình TransNetR dựa

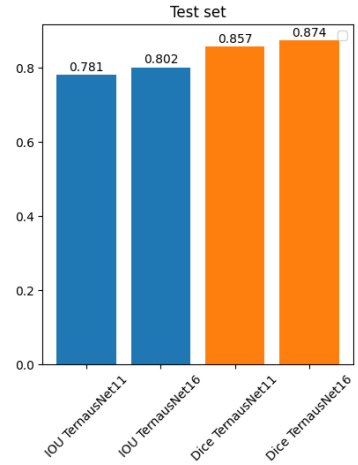

Fig. 10. So sánh kết quả của TernausNet11 và TernausNet16 trên Kvasir

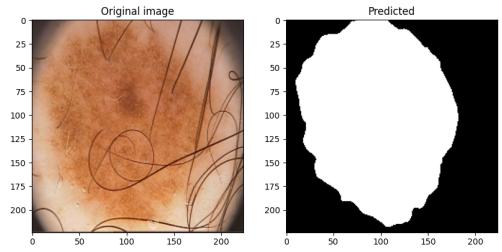

Fig. 11. Kết quả mô hình trên PH2

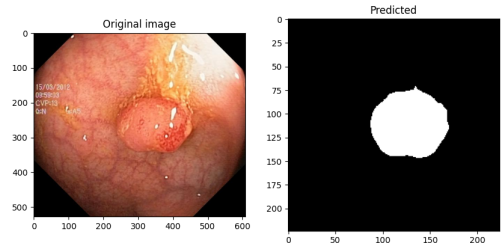

Fig. 12. Kết quả mô hình trên Kvasir

trên kiến trúc Transformer.

VI. Conclusion

Bài nghiên cứu tập trung vào đánh giá hiệu suất của mô hình được đề xuất thông qua quá trình huấn luyện và so sánh với các công trình liên quan trong lĩnh vực phân đoạn hình ảnh y học.

Kết quả thử nghiệm đã cho thấy việc khởi tạo mô hình U-net bằng bộ trọng số của mô hình huấn luyện trước trên tập dữ liệu lớn đã giúp cho việc tinh chỉnh trên tác vụ phân đoạn ảnh y khoa vốn khan hiếm dữ liệu đạt được độ chính xác tốt.

Phần cải tiến của mô hình cũng cho thấy rằng việc thay thế bộ mã hoá bằng mộ bộ mã hoá có mô hình sâu hơn có thể cải thiện khả năng rút trích đặc trưng của mô hình từ đó tăng cao độ chính xác.

Acknowledgment

Chúng em xin bày tỏ lòng biết ơn sâu sắc đến các giảng viên tại Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM đã hỗ trợ và cung cấp những kiến thức quý báu trong quá trình nghiên cứu và viết bài báo này. Chúng em cũng muốn cảm ơn các chuyên gia và bác sĩ đã chia sẻ dữ liệu và kinh nghiệm quý giá, giúp chúng tôi hiểu sâu hơn về lĩnh vực phân đoạn hình ảnh y khoa.

Cuối cùng, không thể không nhắc đến sự đóng góp của các thành viên trong nhóm nghiên cứu, những người đã làm việc không mệt mỏi và đầy đam mê để đạt được kết quả nghiên cứu này.

Trân trọng,

Lê Văn Tấn, Lê Phạm Hoàng Trung, Nguyễn Quang Vinh.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proc. the IEEE Conf. Comput. Vis. Pattern Recognitit. (CVPR)*, 2015.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, 2015.

[3] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, 2019.

[4] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv*, 2021.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv*, 2014.

[6] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv*, 2014.

[7] A. A. Kalinin, V. I. Iglovikov, A. Rakhlin, and A. A. Shvets, "Medical image segmentation using deep neural networks with pre-trained encoders," *Advances in Intelligent Systems and Computing (AISC, volume 1098)*, 2020.

[8] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv*, 2018.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2014.