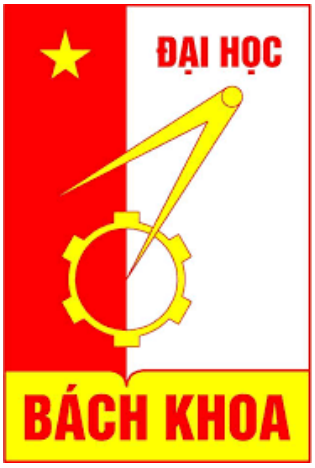


SEMI-UNSUPERVISED SEMENTIC SEGMENTATION

Nhóm 5:

Nguyễn Trung Nghĩa

Bùi Duy Cường

Ngọ Doãn Nghĩa

Lê Văn Tuấn

Võ Đình Thái

Hà Nội, 11/2021

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

TÓM TẮT BÁO CÁO

Cùng với thành tựu đáng kể của học máy trong thị giác máy tính những năm gần đây, rất nhiều mô hình deep learning cho Segmentation hình ảnh y tế đã được công bố, với kết quả ấn tượng. Tuy nhiên, phần lớn các mô hình đó chỉ sử dụng kỹ thuật supervised trong khi các mô hình khác khi sử dụng kỹ thuật như semi-supervised và unsupervised, thu được kết quả không tốt bằng kỹ thuật supervised.

Thông qua một số nghiên cứu đã được công bố cho thấy sự hiệu quả của phương thức Mumford-Shah ở kỹ thuật unsupervised và model Active Contour cho các tác vụ supervised trong bài toán segmentation. Vì vậy, trong báo cáo này, nhóm đã tham khảo và đưa ra một hàm loss kết hợp giữa hai phương pháp trên với một số thay đổi và mở rộng với trường hợp multiphase segmentation. Nó cho phép mô hình deep learning phân đoạn đa lớp với độ chính xác cao hơn thay vì segmentation 2 lớp.

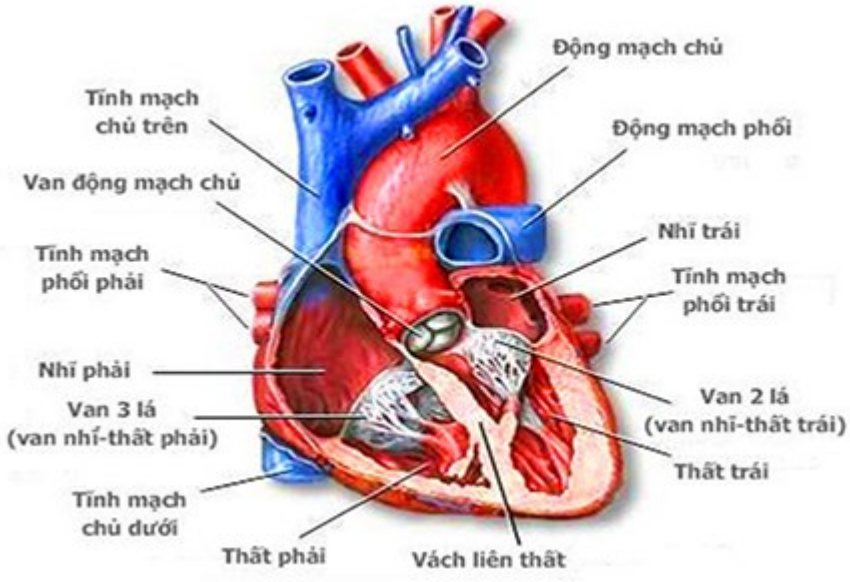
Phương pháp tiếp cận này sẽ được nhóm áp dụng để segmentation tâm thất trái từ hình ảnh MRI tim mạch, trong đó cả màng trong tim và màng ngoài tim được segmentation đồng thời. Phương pháp được đánh giá trên bộ dữ liệu "Automated Cardiac Diagnosis Challenge - ACDCA 2017". Kết quả cho thấy rằng hàm loss đề xuất đạt được kết quả khả quan với độ đo Dice coefficient và Jaccard index. Điều này làm cho thấy sự hiệu quả trong việc segmentation nhiều lớp cho hình ảnh y tế.

.

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN

1.1 Giải phẫu tim và kỹ thuật chụp ảnh tim mạch

1.1.1 Giải phẫu tim



Hình 1.1 Hình ảnh giải phẫu tim

Tim nằm ở trung thất, mặt trên cơ hoành. Tim gồm 4 buồng, nằm trên là 2 buồng nhĩ, nằm dưới là 2 buồng thất. Vách gian thất và vách gian nhĩ phân tim theo trục dọc. Bên ngoài cơ tim được bao bọc bởi màng ngoài tim và bên trong được lót bởi màng trong tim. Hai động mạch vành xuất phát từ gốc động mạch chủ có nhiệm vụ cung cấp máu nuôi dưỡng tim.

- Tâm nhĩ của tim có thành mỏng hơn tâm thất, nằm ở đáy tim. Tâm nhĩ phải có tĩnh mạch chủ từ dưới đổ vào, tâm nhĩ trái có 4 tĩnh mạch từ phổi đổ vào;
- Tâm thất, nhất là tâm thất trái có thành cơ dày hơn tâm nhĩ. Trong tâm thất có các cơ nhú, đầu tự do các cơ có các thừng gân nối vào các lá của van nhĩ thất.

1.1.2 Kỹ thuật chụp cộng hưởng từ tim mạch

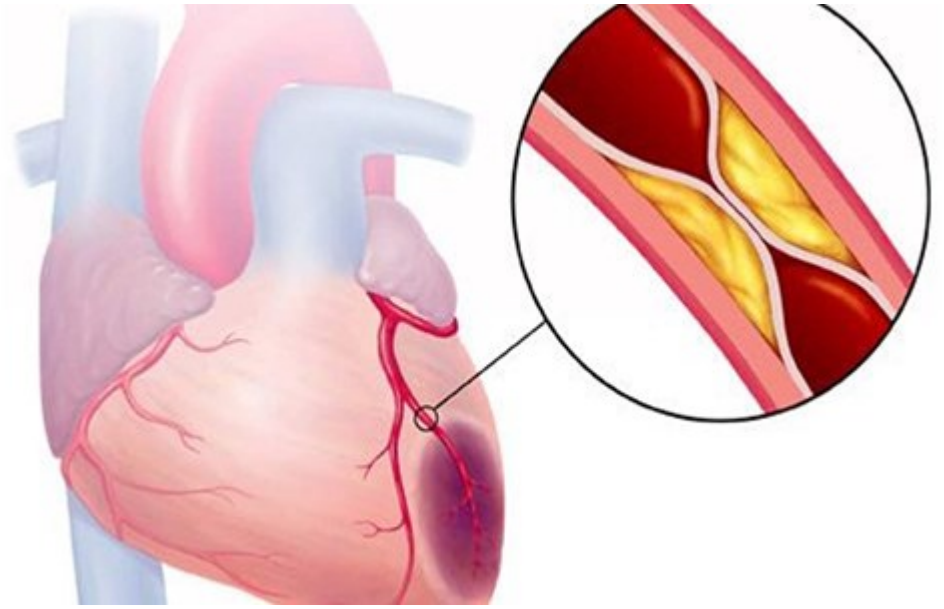
Chụp cộng hưởng từ (MRI) là phương pháp chẩn đoán hình ảnh hiện đại, được sử dụng để kiểm tra hầu hết các cơ quan trong cơ thể, đặc biệt có giá trị trong chụp ảnh chi tiết não hoặc thần kinh cột sống. Kỹ thuật chẩn đoán hình ảnh này sử dụng từ trường mạnh và sóng vô tuyến radio. Nguyên lý của chụp MRI dựa trên tác động của các sóng do máy MRI sản sinh ra, khiến cho các mô trong cơ thể hấp thụ và phóng

thích năng lượng. Sau đó, năng lượng sẽ thu lại, xử lý và chuyển hóa thành các tín hiệu hình ảnh.

Chụp MRI cho hình ảnh có độ phân giải cao và độ tương phản tốt, giúp bác sĩ đánh giá chi tiết tình trạng tổn thương các cơ quan và chẩn đoán bệnh chính xác. Bên cạnh đó, hình ảnh từ MRI có khả năng tái tạo 3D, hỗ trợ tích cực cho việc chẩn đoán bệnh, đưa ra phác đồ điều trị hiệu quả. Trong nhiều trường hợp, chụp MRI tốt hơn so với các kỹ thuật chẩn đoán hình ảnh khác như siêu âm, X-quang hay chụp CT.

Ngoài ra, quá trình chụp bằng MRI không gây tác dụng phụ như chụp X-quang hay chụp CT. Đồng thời, MRI còn cho phép phát hiện các dấu hiệu bất thường ẩn sau các lớp xương mà các phương pháp chẩn đoán hình ảnh khác khó phát hiện được. Trong đó, Cộng hưởng từ tim mạch (cine-MRI) được xem là phương pháp chẩn đoán và theo dõi bệnh lý tim mạch hiện đại nhất hiện nay. Nhờ có kỹ thuật này, các bác sĩ có thể tiên lượng, định hướng điều trị hiệu quả cho người mắc các bệnh về tim mạch để ngăn chặn được diễn tiến bệnh và đưa ra biện pháp chủ động chăm sóc, bảo vệ sức khỏe tốt hơn. Chụp MRI tim cho kết quả nhanh chóng và chính xác hơn X-quang trong chẩn đoán các bệnh về tim mạch.

Cộng hưởng từ tim mạch là kỹ thuật chẩn đoán hình ảnh được chỉ định cho các bệnh nhân mắc hoặc nghi ngờ mắc các bệnh lý về tim. Siêu âm tim là phương pháp hình ảnh đầu tiên. Tuy nhiên, MRI tim đang đóng vai trò ngày càng quan trong trong chẩn đoán, theo dõi và điều trị bệnh lý tim mạch. Kỹ thuật này được ứng dụng rộng rãi trong các bệnh lý mạch vành, bệnh tim bẩm sinh, bệnh cơ tim, van tim...



Hình 1.2 Chụp cộng hưởng từ động mạch vành

Ưu điểm của **chụp MRI tim mạch:**

- Độ phân giải không gian cao
- Mặt phẳng khảo sát không giới hạn
- An toàn, không xâm lấn
- Xác định được đặc tính mô

Tuy nhiên MRI tim bị hạn chế trong một số trường hợp sau:

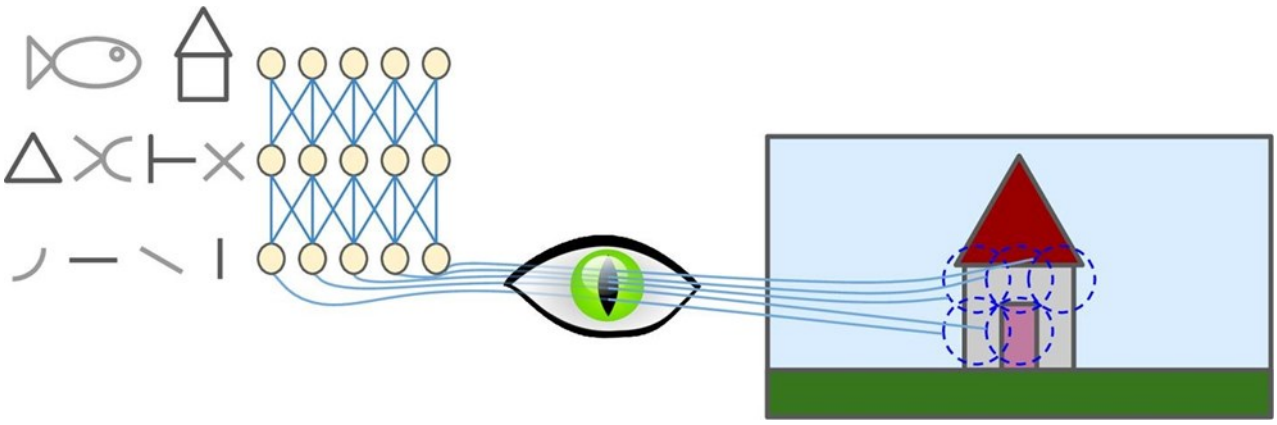
- Chuyển động: tim, lồng ngực, cơ hoành... gây nhiễu ảnh
- Các cấu trúc kích thước nhỏ: <1mm (Ví dụ: Động mạch vành): khó khảo sát

1.2 Deep Computer Vision trong xử lý ảnh.

CNNs xuất hiện từ cuộc nghiên cứu thị giác vỏ não và đã được sử dụng trong bài toán nhận diện ảnh từ năm 1980. Trong những năm gần đây, nhờ có sự phát triển của công nghệ, đặc biệt trong việc xử lý tính toán, số lượng dữ liệu có thể dùng cho việc huấn luyện và các kỹ thuật dùng để huấn luyện các mạng sâu, CNNs đã có thể thực hiện được nhiều bài toán phức tạp. CNNs hỗ trợ cho dịch vụ tìm kiếm hình ảnh, ô tô tự lái, hệ thống tự động phân loại video,… Hơn thế nữa, CNNs không bị hạn chế với các bài toán thị giác, đồng thời được dùng với cả nhận diện giọng nói, xử lý ngôn ngữ tự nhiên.

Vào năm 1958 và 1959, David H. Hubel và Torsten Wiesel thực hiện một chuỗi các thí nghiệm với mèo (vài năm sau đó là với khỉ), đưa ra được những đặc điểm quan

trọng trong cấu trúc của vỏ não thị giác. Trong đó, họ đưa ra kết luận rằng nhiều nơ-ron trong vỏ não thị giác có một vùng tiếp nhận cục bộ (local receptive field), nghĩa là chúng chỉ phản ứng với những kích thích nhìn thấy được ở một vùng bị giới hạn của vùng nhìn thấy (như trong hình 1.3, các vùng tiếp nhận cục bộ được biểu diễn bằng các vòng tròn nét đứt). Các vùng tiếp nhận của các nơ-ron khác nhau có thể chồng lấn lên nhau và giao của tất cả những vùng đó tạo nên vùng nhìn thấy chung Hình 1.3, vì các tín hiệu thị giác tạo liên kết liên tiếp qua não, nơ-ron phản ứng với nhiều nội dung phức tạp khi các trường tiếp nhận lớn.



Hình 1.3 Nơ-ron sinh học với vùng vỏ não thị giác phản ứng với vật thể cụ thể bằng những vùng nhìn thấy nhỏ gọi là các vùng tiếp nhận

Hơn thế nữa, hai tác giả kết luận rằng một vài nơ-ron chỉ phản ứng với những ảnh trên đường nằm ngang trong khi một vài tế bào khác chỉ phản ứng với những đường thẳng với những hướng khác nhau (hai nơ-ron có thể có vùng tiếp nhận giống nhau nhưng có thể phản ứng với các đường có hướng khác nhau). Họ cũng đề cập rằng một vài nơ-ron có vùng tiếp nhận lớn và chúng có thể phản ứng với những chi tiết phức tạp hơn, những chi tiết là sự kết hợp của những chi tiết mức thấp hơn. Kết quả này dẫn đến ý tưởng rằng những nơ-ron cấp cao được dựa trên đầu ra của lân cận các nơ-ron cấp thấp hơn (Hình 3, chú ý rằng mỗi nơ-ron được kết nối chỉ với vài nơ-ron lớp trước). Cấu trúc quan trọng này có thể dùng để phát hiện tất cả các loại cấu trúc phức tạp ở bất cứ khu vực nào của trường thị giác.

Ngày nay, với sự phát triển của công nghệ, sự nâng cấp mạnh mẽ của phần cứng giúp tăng tốc độ tính toán rất mạnh mẽ thì các bài toán như nhận dạng, hay phân đoạn

hình ảnh trước đây đã sử dụng mạng neuron ngày càng được nghiên cứu, và áp dụng nhiều. Kết quả thu được là rất tốt khi so với các phương pháp xử lý ảnh truyền thống trước đây.

1.3 Đặt vấn đề

Segmentation hình ảnh là một trong những khía cạnh, ứng dụng quan trọng bậc nhất đối với thị giác máy tính [1]. Với hình ảnh đã được segmentation, người ta có thể có thêm thông tin về đối tượng quan tâm cho các bước phân tích ảnh tiếp theo như trích xuất hay nhận diện đối tượng. Nói một cách tổng quát, trong quá trình segmentation, mỗi pixel của một hình ảnh được phân loại thành một lớp nhất định, vì vậy nó có thể được coi là bài toán phân loại dựa trên pixel.

Trong lĩnh vực y tế, automatic image segmentation có tầm quan trọng nhất định, vì trên thực tế là các phương pháp thủ công truyền thống không chỉ tốn kém thời gian mà còn đòi hỏi kiến thức chuyên môn cao. Bên cạnh đó, các hình ảnh y tế thường bị hạn chế, không có nhãn và các lớp thường mất cân bằng [2]. Trong segmentation ảnh cổ điển, đặc biệt là đối với ảnh y tế, phương thức Mumford-Shah từ công trình nghiên cứu tiên phong của Mumford và Shah [3] đã là chủ đề trọng tâm được đưa ra bàn luận trong suốt hai thập kỷ qua. Từ đó, một loạt các phương pháp segmentation đã được phát triển như active contour model (ACM) và phương pháp lấy ngưỡng[4], phương pháp xấp xỉ [5]. Mặc dù có những ưu điểm như cho các đường bao khép kín và độ chính xác theo subpixel, active contour, tuy nhiên các phương pháp lấy ngưỡng có nhược điểm đó là người ta cần khởi tạo các giá trị, ngưỡng thủ công cho đường bao. Do đó gây khó khăn cho các bài toán segmentation tự động. Từ đó, nhu cầu về các phương pháp đáng tin cậy và chính xác hơn, đồng thời có thể tự động segmentation các hình ảnh y tế ngày càng được quan tâm.

Gần đây, hiệu quả đã được chứng minh của deep learning (DL) trong segmentation hình ảnh đã trở thành tiền đề quan trọng cho các phương pháp segmentation tự động ngày nay để áp dụng đối với hình ảnh y tế. Có rất nhiều mô hình DL, đặc biệt là convolution neural network (CNN) đã được giới thiệu và nhận được nhiều kết quả đáng chú ý [6–8), trong đó không thể không kể đến CNN dựa trên kiến trúc U-net [9] với một số cải tiến trong model, metrics và hàm loss. Trong quá trình huấn luyện mô

hình CNN, thông qua các lần so sánh các mask dự đoán và mask đã được segmentation bởi bác sĩ, hàm loss được tối thiểu hóa bằng cách cập nhật các tham số mô hình thông qua các phương pháp gradient descent. Do vậy, việc định nghĩa hàm loss chuẩn xác sẽ đóng góp đáng kể vào việc tối ưu hóa mô hình. Đối với segmentation hình ảnh y tế, Dice coefficient (DC), cross-entropy (CE) hoặc Tversky loss thường được ưu thích sử dụng hơn cả [9–13]. Tuy nhiên, những metric đó thường thiếu các ràng buộc ở vùng đường bao, dẫn đến các kết quả không mong muốn ở gần ranh giới khoanh vùng [14]. Một số nhà nghiên cứu đã chỉ ra rằng hiệu quả của U-Net có thể được nâng cao bằng cách thiết kế các hàm loss khác nhau [14, 15].

Mặt khác, ý tưởng kết hợp mô hình active contour và phương pháp DL để segmentation hình ảnh hiệu quả [8, 10] đã trở nên phổ biến trong năm qua. Điều này là do phương pháp active contour dựa trên ngưỡng ưa chuộng hơn vì nó cho phép đường cong thay đổi cấu trúc liên kết của nó trong quá trình segmentation. Tuy nhiên, khả năng segmentation của nó phụ thuộc đáng kể vào việc khởi tạo đường cong. Việc khởi tạo không chuẩn xác có thể khiến mô hình bị mắc kẹt ở điểm cực tiểu cục bộ. Điều này có thể được xử lý bằng cách sử dụng phương pháp DL như một cách segmantation thô trước khi tinh chỉnh lại đường bao bằng cách sử dụng ACM. Trong một phương án khác, các active contour lấy ý tưởng từ phương thức Mumford-Shah và biến thể của nó đã được sử dụng làm các hàm loss trong quá trình huấn luyện mạng nơ-ron [16, 17]. Trong nghiên cứu gần đây, Kim và Ye [18] đã đề xuất một cách tiếp cận mới để kết hợp đầu ra softmax của mô hình DL với phương thức Mumford-Shah. Tận dụng lợi thế của các phương pháp tiếp cận đã đề cập ở trên, trong dự án này, nhóm sử dụng phương thức Active Contour được để multiphase segmentation sao cho cả hai vùng màng trong tim và màng ngoài tim được segmentation đồng thời. Nhóm cũng lập một bảng kết quả để so sánh trong trường hợp có và không có sử dụng phương thức Mumford-Shah. Những đóng góp chính của nhóm trong dự án này bao gồm:

1. Giới thiệu một hàm loss mới kết hợp phương thức Mumford-Shah và Active Contour;
2. Xây dựng mô hình cuối end-to-end để multiphase segmentation, được huấn luyện với dữ liệu hạn chế.

3. Đạt được kết quả đầy hứa hẹn trên cơ sở dữ liệu ACDCA 2017 đồng thời giảm thời gian huấn luyện.

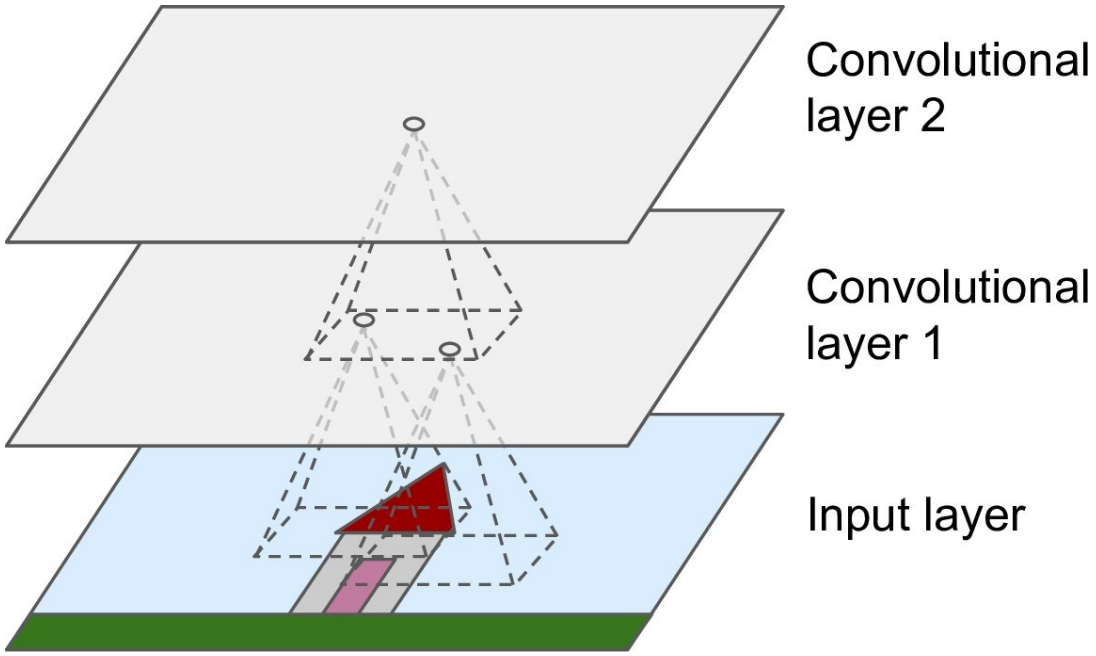
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Deep Computer Vision với CNNs

Những nghiên cứu về vùng vỏ não thị giác là nguồn cảm hứng tạo nên Neocognitron (giới thiệu vào năm 1980) mà hiện nay được gọi là Mạng thần kinh tích chập (Convolutional Neural Networks). Một cột mốc quan trọng là một nghiên cứu vào năm 1998 bởi YannLeCun et al, giới thiệu cấu trúc nổi tiếng LeNet-5, đã được các ngân hàng đưa vào sử dụng để phát hiện chữ viết tay và kiểm tra số. Cấu trúc này gồm vài khối như các lớp Fully Connected và hàm Activation Sigmoid. Tuy nhiên, nó cũng giới thiệu hai khối mới là Convolutional và pooling layer.

2.1.1 Lớp Convolutional

Block quan trọng nhất của mạng CNN là lớp convolutional: nơ-ron ở lớp convolutional đầu tiên không được kết nối tới từng pixel trong ảnh mà chỉ những pixel trong vùng tiếp nhận (hình 2.1). Đổi lại, mỗi nơ-ron ở lớp convolutional thứ hai chỉ được kết nối với nơ-ron nằm trong một hình chữ nhật nhỏ ở lớp đầu tiên. Cấu trúc này cho phép mạng lưới kết hợp vào các đặc trưng nhỏ mức thấp ở lớp hidden đầu tiên, sau đó gộp chúng vào các đặc trưng lớn ở mức cao hơn tại lớp hidden tiếp theo,… Cấu trúc có thứ bậc này phổ biến với ảnh thực, là một trong những lý do CNN đem lại hiệu quả tốt với bài toán nhận diện ảnh.

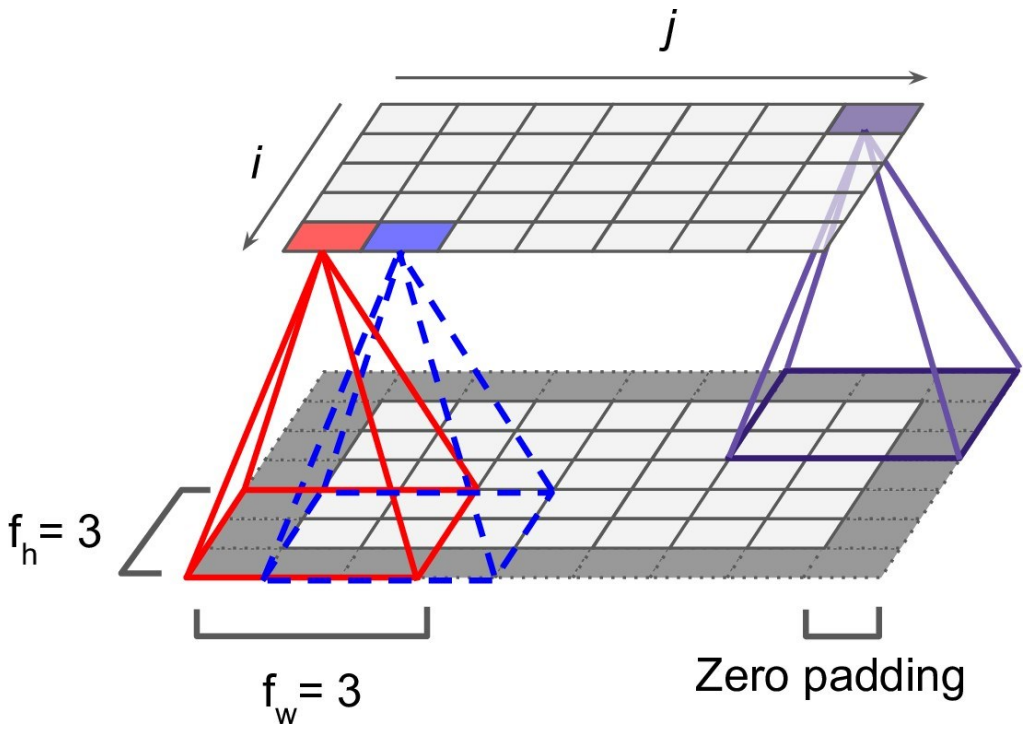


Hình 2.1 Các lớp CNN với trường tiếp nhận cục bộ hình chữ nhật

Một nơ-ron được biểu thị hàng i, cột j của một lớp cho trước được kết nối với đầu ra của các nơ-ron ở lớp trước nằm ở hàng i tới $i + fh - 1$, cột j tới $j + fw - 1$, trong đó fh và fw là chiều dài và rộng của trường quan sát. Để một lớp có cùng chiều dài và rộng như lớp trước, thường thêm viền 0 vào đầu vào, gọi là *zero padding.*

Padding ảnh là việc thêm các pixel vào các cạnh để mở rộng ảnh. Có ba kiểu padding thường thấy là: duplication padding, zero padding và replication padding.

- Replication padding là lặp lại các giá trị viền của bốn hàng pixel tương ứng.
- Duplication padding là các giá trị của các hàng padding tương ứng sẽ là giá trị của hàng đối diện. Ví dụ padding của hàng 1 sẽ là giá trị của hàng 5, trường hợp ma trận có 5 hàng.
- Zero padding là viền bên ngoài gồm 1 hàng chỉ mang giá trị 0.



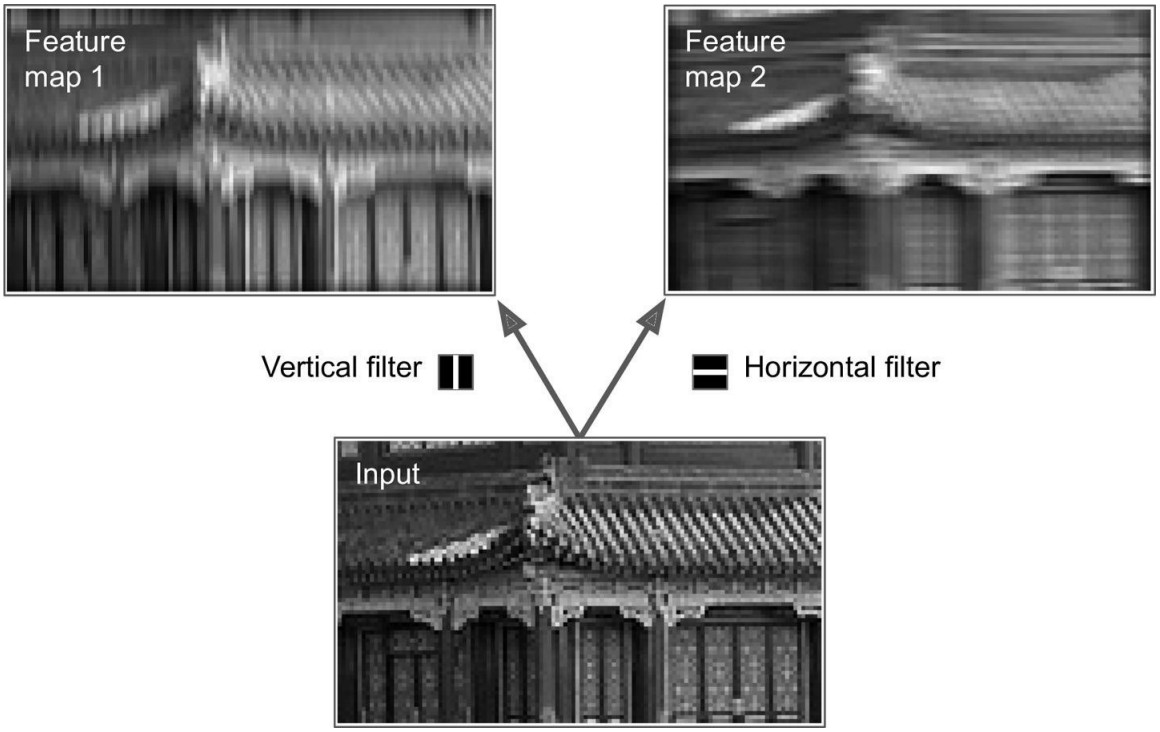
Hình 2.2 Liên kết giữa các lớp và Zero padding

❖ **Feature map**

Các trọng số của một nơ-ron có thể được hiển thị như một bức ảnh nhỏ có kích thước của trường tiếp nhận. Ví dụ hình 2.3 cho hai tập trọng số gọi là filters (hoặc convolutional kernels). Tập trọng số đầu (filter đầu) được biểu diễn dưới dạng một ma trận vuông màu đen với một đường màu trắng thẳng đứng chính giữa (đó là một ma trận 7x7, hầu hết các pixel có giá trị 0 ngoại trừ một cột trung tâm chính giữa mang giá trị 1); nơ-ron dùng trọng số này sẽ bỏ qua mọi thứ trong trường quan sát ngoại trừ đường thẳng chính giữa (vì tất cả đầu vào sẽ được nhân với 0, ngoại trừ đường thẳng ở vị trí chính giữa). Filter thứ hai là một ma trận vuông đen với đường trắng nằm ngang ở chính giữa. Một lần nữa, nơ-ron dùng các trọng số này để bỏ qua mọi thứ trong trường quan sát (Receptive Field) ngoại trừ đường ngang chính giữa.

Nếu tất cả nơ-ron trong một lớp dùng cùng một loại filter dọc (và cùng bias), đầu vào của mạng là ảnh trong Hình 2.3 (ảnh ở giữa phía dưới), đầu ra của ảnh sẽ là ảnh bên góc trái phía trên. Chú ý rằng những đường thẳng dọc màu trắng được tăng cường khi phần còn lại bị làm mờ. Tương tự, bức ảnh bên phải thu được khi nơ-ron dùng cùng filter đường ngang, những đường ngang cũng được tăng cường khi các vùng

xung quanh bị mờ. Do đó, một lớp có các nơ-ron dùng cùng một loại filter sẽ tạo thành một feature-map.



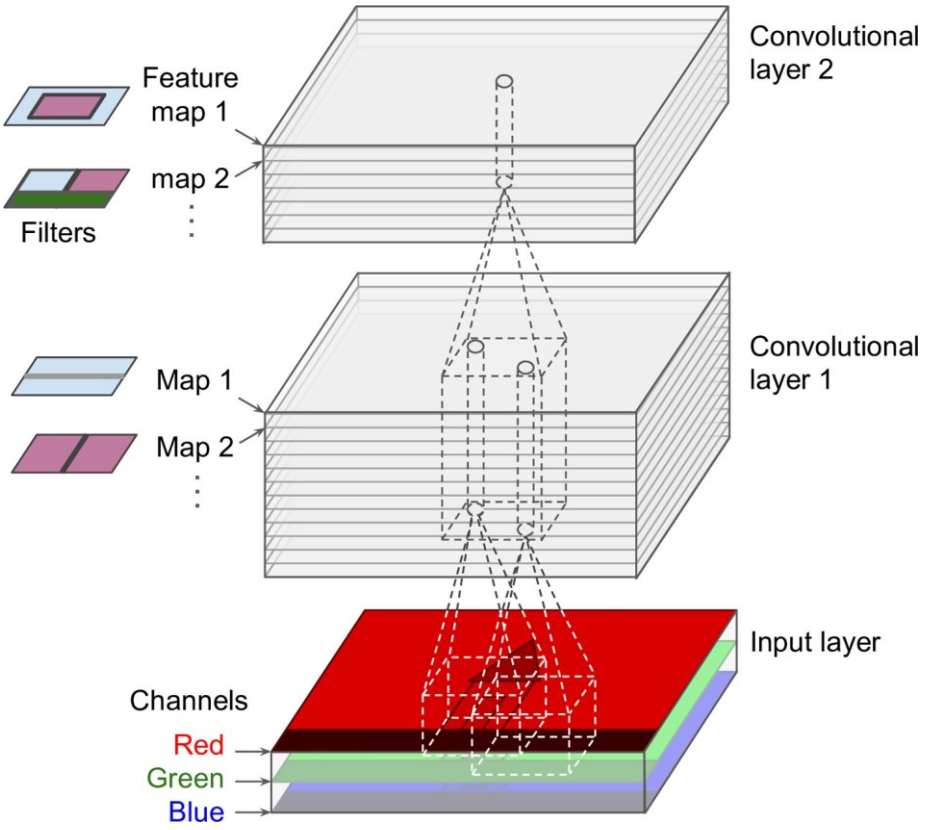
Hình 2.3 Dùng hai filter khác nhau thu được hai feature-map

Trong thực tế, một lớp convolutional có nhiều filter và đầu ra sẽ đưa một feature map với mỗi filter. Hình 2.4 mỗi nơ-ron tương ứng với một pixel ở mỗi feature map và tất cả nơ-ron trong một feature map có cùng tham số (nghĩa là cùng weight và bias). Nơ- ron ở các feature map khác nhau dùng những tham số khác nhau. Một nơ-ron của receptive field như đã mô tả nhưng được mở rộng qua tất cả các lớp trước của feature map. Tóm lại, một lớp convolutional đồng thời dùng nhiều filter được huấn luyện thành đầu vào, khiến nó có khả năng phát hiện nhiều loại đặc trưng ở bất cứ đâu của đầu vào.

Ảnh đầu vào cũng được tạo bởi các lớp nhỏ, mỗi lớp là một kênh màu. Có ba kênh cơ bản: đỏ, lục,lam (RGB). Ảnh xám chỉ có một kênh màu nhưng vài ảnh có thể có nhiều hơn, ví dụ như ảnh vệ tinh thu các tần số khác nhau của ánh sáng như hồng ngoại.

Giả thiết, một nơ-ron ở hàng i, cột j của feature map k tại một lớp convolutional l được kết nối với đầu ra của các nơ-ron ở lớp *l-1,* nằm ở hàng i x $sh + fh - 1$ và cột j x fw tới j x $sw + fw - 1$, dựa vào các feature map ở lớp *l-1*. Chú ý rằng tất cả nơ-ron

nằm ở cùng hàng *i* và cột *j* nhưng ở những feature map khác nhau đều được kết nối vào đầu ra của chính xác nơ-ron tương tự của lớp trước.



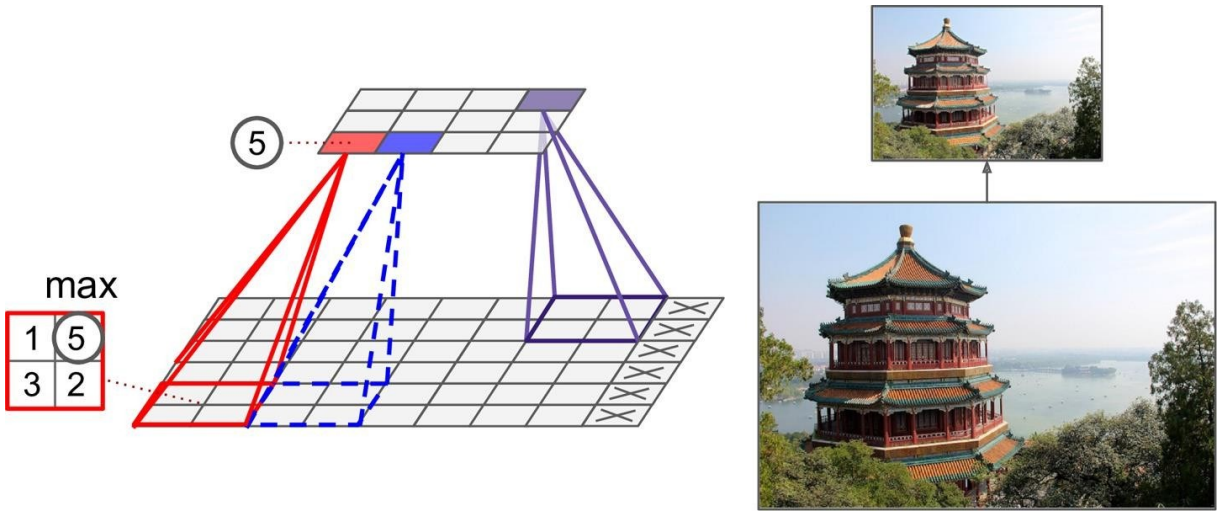
Hình 2.4 Lớp Convolutional với nhiều feature map và ảnh đầu vào có ba kênh

2.1.2 Pooling layer

Mục đích của lớp Pooling là *subsample* (co ảnh) ảnh đầu vào để giảm độ phức tạp khi tính toán, giảm thiểu số tham số (giảm khả năng overfitting) và dung lượng bộ nhớ.

Cũng như lớp convolutional, mỗi nơ-ron ở lớp Pooling được kết nối với đầu ra của một số nơ-ron ở lớp trước nằm trong trường quan sát chữ nhật nhỏ. Một pooling nơ-ron không có các trọng số (weights), nó chỉ thu gọn đầu vào dùng một hàm thu lại là max hoặc trung bình. Hình 2.5 thể hiện một lớp max-pooling (dạng phổ biến nhất cho lớp pooling). Ở ví dụ này, dùng kernel pooling có kích thước 2x2, stride bằng 2 và không padding. Chỉ giá trị lớn nhất của đầu vào ở mỗi trường quan sát được dùng cho lớp tiếp theo trong khi bỏ qua những giá trị khác.

Ví dụ ở hàng thấp nhất bên trái của trường quan sát trong hình 2.5, giá trị đầu vào là 1, 5, 3, 2, do đó, giá trị lớn nhất là 5 được truyền tới lớp tiếp theo. Vì stride là 2, ảnh đầu ra có chiều dài và rộng chỉ bằng nửa ảnh ban đầu (không padding).

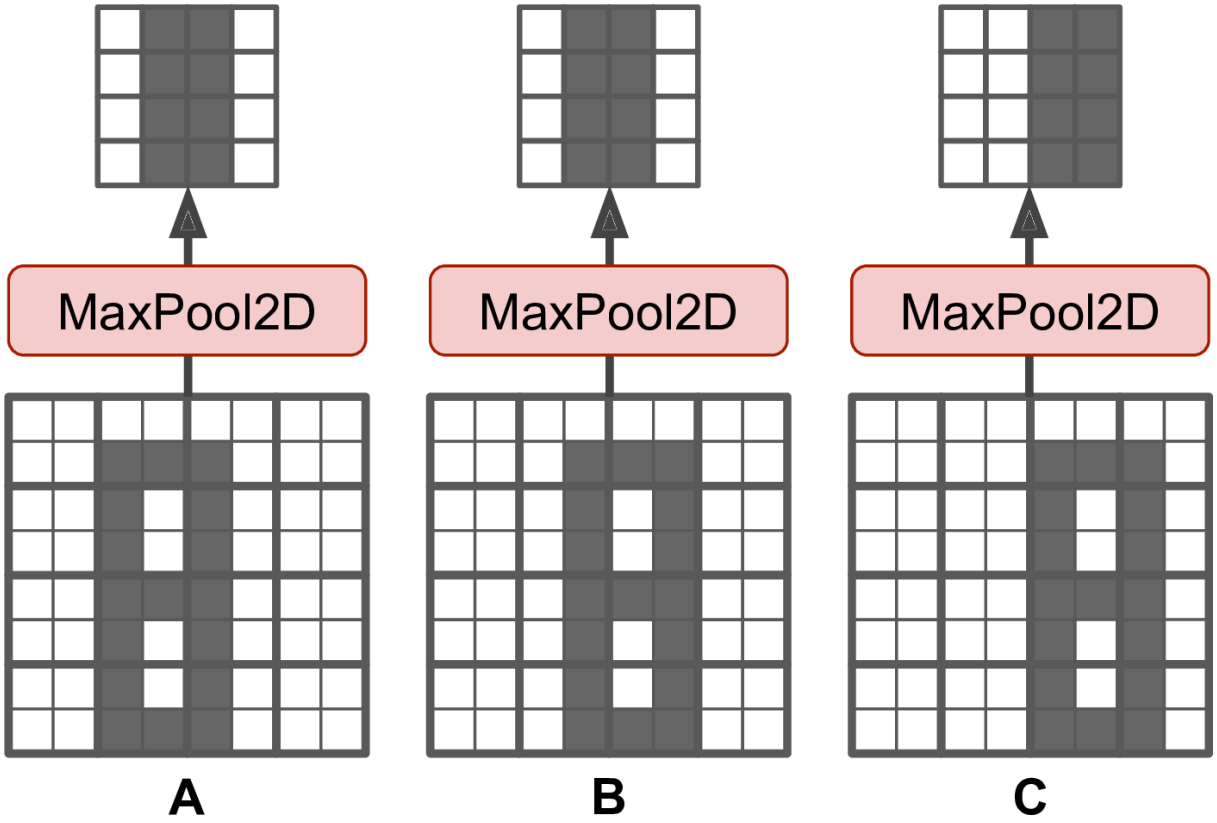


Hình 2.5 Max pooling layer (2x2 pooling kernel, stride 2, no padding)

Ngoài việc giảm thiểu tính toán, tham số và dung lượng bộ nhớ sử dụng, một lớp max pooling được đánh giá là "bất biến" với những thay đổi nhỏ như hình 2.6. Giả thiết rằng những pixel sáng có giá trị thấp hơn những pixel tối và chọn ba ảnh A, B, C qua lớp max pooling với kernel size là 2x2 và stride là 2. Ba ảnh tương tự nhau nhưng ảnh B dịch một pixel về bên phải và ảnh C là hai pixel. Có thể thấy đầu ra của lớp max pooling cho ảnh A và B tương tự nhau, đây chính là "bất biến" với thay đổi. Với ảnh C, đầu ra bị dịch 1 pixel về phía bên phải nhưng vẫn có 75% là bất biến. Việc dùng lớp max pooling ở mỗi lớp của CNN có thể đạt được vài bất biến với thay đổi với những tỷ lệ lớn. Hơn thế nữa, max pooling cần một lượng dịch chuyển bất biến nhỏ và một tỷ lệ bất biến nhẹ. Bất biến (dù có thể bị giới hạn) có thể thành công trong trường hợp khi kết quả dự đoán không phụ thuộc vào chi tiết như những bài toán phân loại.

Tuy nhiên, max pooling cũng có một vài mặt tối. Đầu tiên là rõ ràng max pooling có tính "phá hoại", dù với kích thước kernel nhỏ là 2x2 và stride là 2, kết quả đầu ra sẽ bị nhỏ đi hai lần cả về chiều rộng và chiều dài (diện tích sẽ giảm bốn lần). Điều này làm mất 75% giá trị đầu vào.

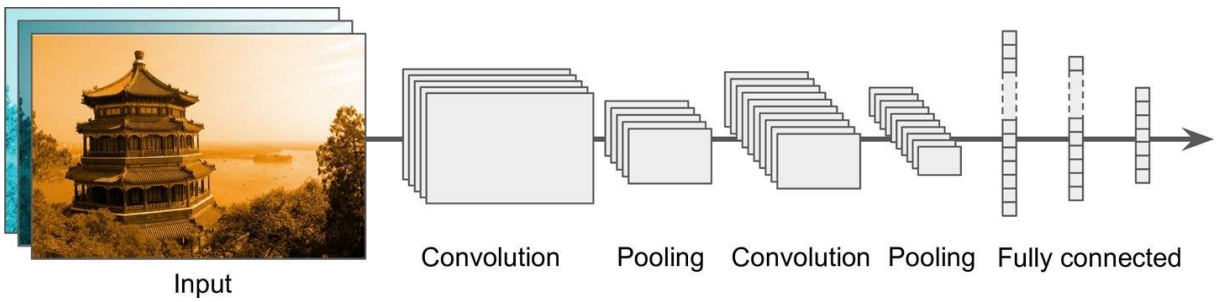
Với vài ứng dụng thì không mong muốn việc bất biến. Bài toán Semantic Segmentation (phân loại từng pixel trong ảnh thuộc vật thể nào), rõ ràng nếu ảnh đầu vào được dịch chuyển chỉ một pixel về phía bên phải, đầu ra phải được dịch chuyển một pixel về phía phải. Mục đích của trường hợp này là cân bằng, không phải bất biến: một thay đổi nhỏ của đầu vào cần dẫn đến một thay đổi nhỏ của đầu ra.



Hình 2.6 Max-pooling bất biến với những thay đổi nhỏ

i. Cấu trúc CNN

Cấu trúc CNN thông thường gồm một vài lớp convolutional, mỗi lớp theo sau bởi một lớp ReLU, một lớp Pooling, tiếp theo là một lớp Convolution (và ReLU),…



Hình 2.7Cấu trúc CNN thông thường

Ảnh ngày càng nhỏ lại nhưng sâu hơn qua mạng (nhiều feature map hơn). Phần đầu của mạng là một ảnh thông thường và lớp đầu ra cuối cùng là dự đoán (lớp Softmax đánh giá xác suất).

2.1.3 Sementic Segmentation

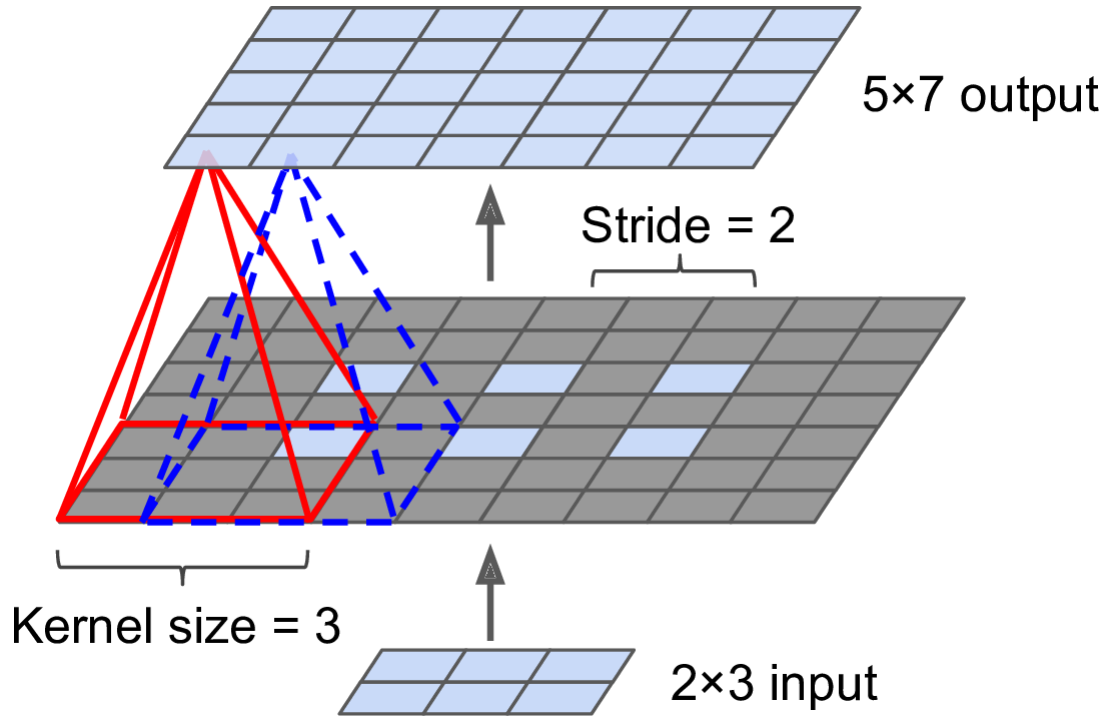
Trong Sementic segmentation, mỗi pixel được phân loại dựa theo các lớp vật thể của nó (đường, tòa nhà, cây, biển báo,…) như ở hình 2.8. Chú ý là những vật thể khác nhau của cùng một lớp thì không phân biệt riêng. Ví dụ tất cả xe đạp ở phía bên phải của ảnh phân đoạn được gộp thành một khối pixel. Khó khăn chính của việc này là khi những bức ảnh qua mạng CNN thông thường sẽ bị mất độ phân giải không gian (do các lớp có strike lớn hơn 1). Vì vậy với mạng CNN thông thường có thể biết được trong ảnh có người nhưng không xác định được chính xác vị trí của người trong ảnh.

Giống như với bài toán phát hiện vật thể, có nhiều cách tiếp cận để xử lý vấn đề này, một số cách khá phức tạp. Tuy nhiên, một giải pháp khá là đơn giản được đề xuất vào năm 2015 bởi Jonathan Long. Tác giả dùng một mạng CNN đã được huấn luyện và cho vào lớp FCN. Mạng CNN dùng stride chung là 32 với ảnh đầu vào, nghĩa là lớp cuối cùng cho ra feature map nhỏ hơn 32 lần so với ảnh đầu vào. Việc này rõ ràng mang đến nhiều sai sót do đó họ thêm một lớp upsampling giúp tăng độ phân giải 32 lần.

Hình 2.8 Semantic segmentation

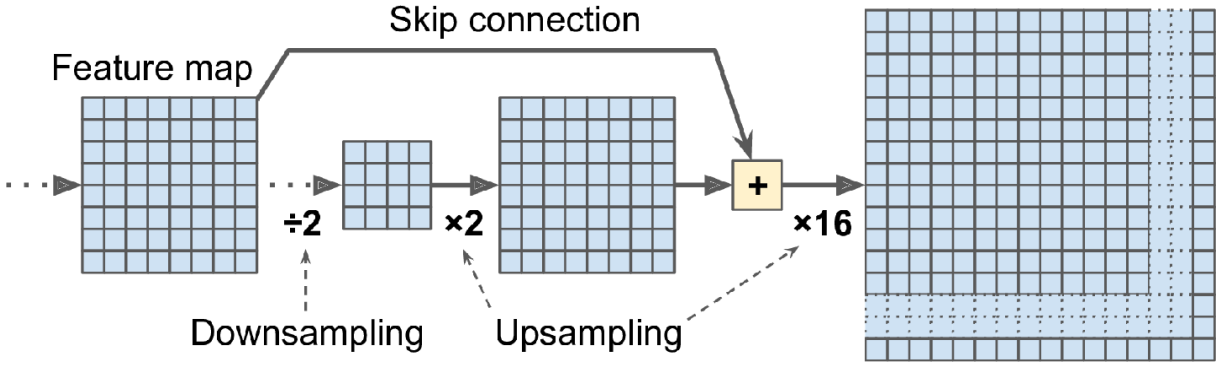
Có vài giải pháp cụ thể cho việc upsampling như bilinear interpolation (nội suy song tuyến tính) nhưng chỉ có thể gấp 4 lần hoặc 8 lần. Thay vào đó họ dùng lớp transposed convolutional layer: tương đương với việc kéo căng ảnh bằng việc chèn thêm các hàng và cột bằng 0, sau đó thực hiện phép convolution thông thường (hình 2.9). Một vài phương pháp chọn dùng lớp convolution với stride rất nhỏ (ví dụ stride = ½ ở Hình 2.9). Lớp transposed convolution có thể được dùng cho vài trường hợp gần với phép nội suy tuyến tính, vì đó là một lớp có thể huấn luyện nên nó sẽ luôn cải thiện

trong quá trình training.



Hình 2.9 Upsampling dùng lớp Transposed convolution

Phương pháp trên vẫn chưa chính xác. Để tăng độ chính xác, tác giả đề xuất thêm một lớp skip connection từ những lớp thấp: ví dụ họ tăng ảnh đầu ra lên 2 lần thay vì 32, sau đó thêm đầu ra của lớp thấp hơn có độ phân giải gấp đôi. Sau đó họ upsample kết quả 16 lần, điều này làm cả ảnh được upsample lên 32 lần (Xem 2.10). Phương pháp này tái tạo độ phân giải không gian đã bị mất ở lớp pooling trước. Ở những cấu trúc tốt nhất, có thể dùng một Skip connection thứ hai để tái tạo những chi tiết tốt hơn từ những lớp thấp hơn: upscale x2, thêm đầu ra của lớp thấp hơn (tỷ lệ thích hợp), upscale x2, thêm đầu ra của lớp thấp hơn nữa, upscale x8.



Hình 2.10 Lớp skip tái tạo độ phân giải không gian từ những lớp thấp hơn

Nó có thể được đưa về tỷ lệ lớn hơn ảnh gốc, phương pháp này có thể dùng để tăng độ phân giải của ảnh, kỹ thuật này gọi là super-resolution.

Ngoài ra còn có instance segmentation. Instance segmentation tương tự như semantic segmentation nhưng không gộp tất cả vật thể vào chung một lớp mà các vật thể chung một lớp được phân chia rõ ràng.

Hiện nay, các mô hình của Deep Computer Vision được phát triển vô cùng nhanh và đều đưa vào CNN. Mục tiếp theo sẽ là mô hình đề xuất cho việc phân đoạn ảnh.

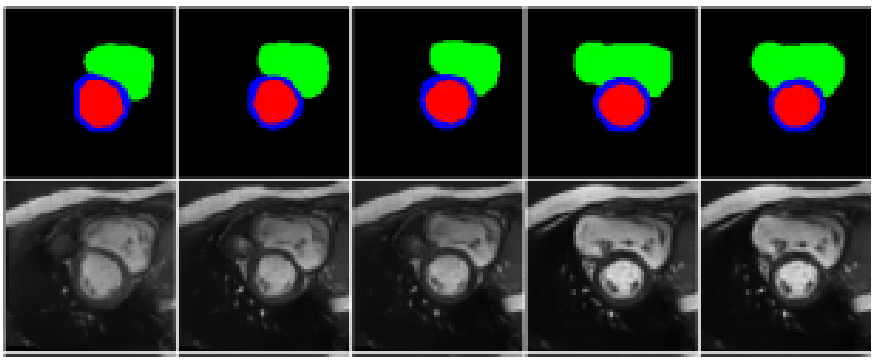
CHƯƠNG 3. GIẢI QUYẾT BÀI TOÁN

3.1 Mô tả dữ liệu

Dữ liệu sử dụng trong dự án này dựa trên cơ sở dữ liệu do hội thảo "Automatic Cardiac Diagnosis Challenge (ACDC)" tổ chức cùng với Hội nghị quốc tế lần thứ 20 về International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), vào ngày 10 tháng 9 năm 2017 ở Thành phố Quebec, Canada [25]. Cơ sở dữ liệu huấn luyện bao gồm:

- Tập ảnh chụp 4D cine-CRI từ 100 bệnh nhân, mỗi ảnh bao gồm segmentation masks cho tâm thất trái (LV), cơ tim (Myo) và tâm thất phải (RV) ở cuối tâm thu (ES) và giai đoạn cuối tâm trương (ED) của mỗi bệnh nhân.
- Tập cine-MRI scans từ 50 bệnh nhân không có segmentation masks phục vụ mục đích kiểm thử trong cuộc thi.

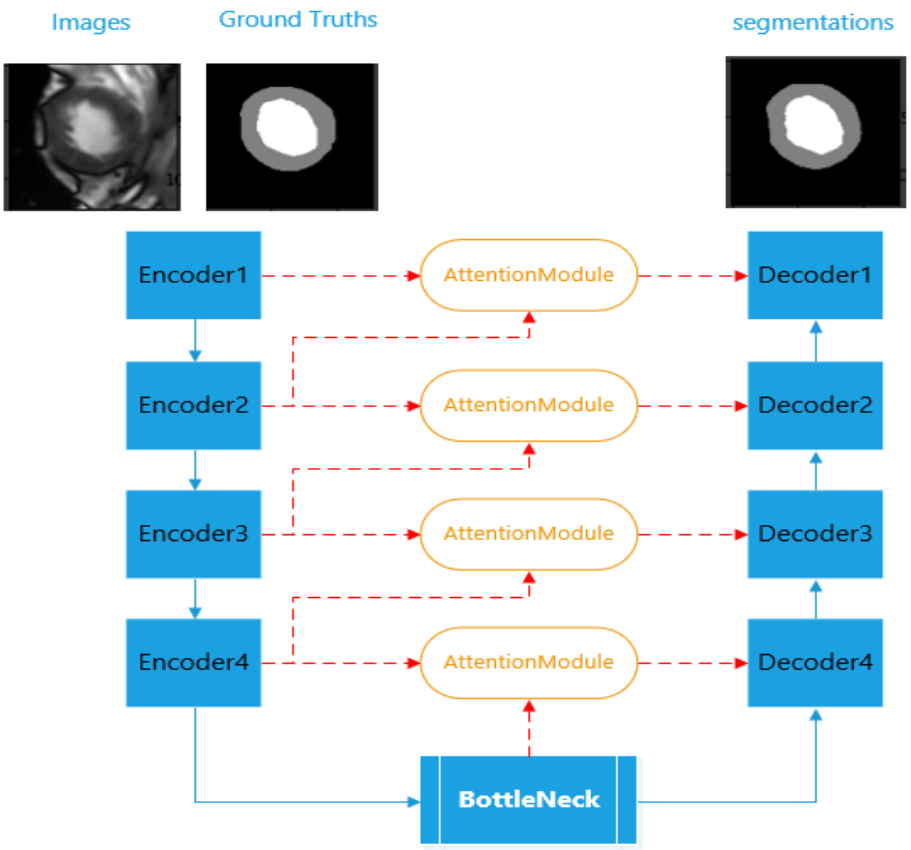
Do cuộc thi đã đóng, nên nhóm quyết định chỉ sử dụng tập dữ liệu từ 100 bệnh nhân có segmentation masks, kết hợp với một số kỹ thuật data augmentation để thực hiện huấn luyện và kiểm thử. Cơ sở dữ liệu này được tách thành tập huấn luyện và tập thử nghiệm với tỷ lệ 8: 2 để đánh giá mô hình segmentation hình ảnh.



Hình 3.1 Ảnh Cine-MRI và các mask tương ứng

3.2 Mô hình đề xuất

U-Net được đề xuất như một trong những mô hình kinh điển để giải quyết bài toán phân đoạn ảnh. Nhóm đã sử dụng U-net [9] làm cơ sở để tùy chỉnh sao cho phù hợp với bài toán. Mô hình bắt nguồn từ cấu trúc U-net với hai nhánh như được thấy trong Hình 3.2.



Hình 3.2 Mô hình đề xuất

Nhiệm vụ chính của bộ encoder là trích xuất thông tin từ hình ảnh đầu vào và chuyển thông tin cần thiết bằng cách skip layers cho decoder. Ban đầu, hình ảnh đầu vào được chuẩn hóa trong khoảng [0, 1] của kênh chiều cao và chiều rộng (với ảnh xám), sau đó được đưa vào khối encoder. Nó được đưa xuống bốn khối để downsample (bốn khối encoder), mỗi khối chứa một lớp tích chập 2D sau đó được batch normalization và hàm activation Swish [21]. Hàm activation Swish được sử dụng vì nó có lợi thế đối với ma trận thưa (ảnh y tế), giống như hàm activation ReLU [22], đồng thời nó không bị ràng buộc ở trên, điều này đảm bảo rằng đầu ra thu được không quá lớn.

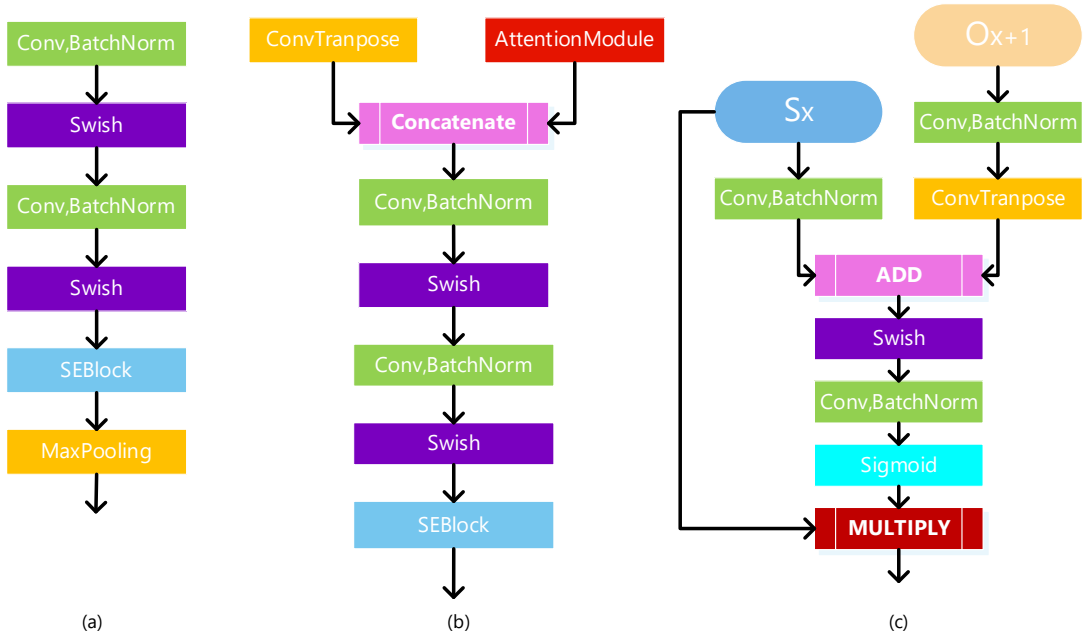
Bằng cách sử dụng skip layer, đầu ra của encoder là bốn skip layer được lấy từ 4 khối downsample trên, được ký hiệu là:

$$S_1, S_2, S_3, S_4;\ S_i \in R^{H_i * W_i * C_{si}}$$

$$H_i = 2\,H_{i+1}$$

$$W_i = 2\,W_{i+1} \text{ với i} = \overline{1,4}$$

H_i, W_i là chiều cao, chiều rộng của feature map. Sau đó, S_1, S_2, S_3, S_4 được cho vào một module attention để học cách thu thập thông tin chính xác hơn trước khi ghép nối với các đầu ra của khối decoder. Việc sử dụng module attention trước khi ghép giúp mạng đặt trọng số lớn vào các features quan trọng trước khi skip layers. Điều này có thể cho phép attention nhiều vào một phần cụ thể của đầu vào. Kết quả là, feature map bỏ qua phần không quan trọng và phân bổ sự chú ý lên các phần có liên quan. Chi tiết cấu trúc của mô hình được minh họa trong Hình 3.3.



Hình 3.3(a) Encoder Block (b) Decoder Block (c) Attention Module

Đối với bộ decoder, nó bao gồm 4 khối upsampling, và khối bootech neck với đầu ra là O_1, O_2, O_3, O_4, O_5 ($O_i \in R^{H_i * W_i * C_{di}}$), với i = $\overline{1,5}$, và đầu ra để so sánh với ground truth là $O_1 \in R^{H*W*1}$. O_{i+1} và S_i được đưa vào attention module, trả về S_i' có cùng kích thước với S_i. Khi đó, S_i' và O_{i+1} được concatenated thành $[O_{i+1}, S_i']$ với i = $\overline{1,4}$, sau đó đi qua mỗi khối upsampling để tạo thành Oi. Mỗi khối upsampling bao gồm một khối Squeeze-and-Excitation (SEBlock), theo sau là khối soft residual được sửa đổi từ khối original residual trong Resnet [23] (bằng cách thay thế tích chập bằng tích chập theo chiều sâu) để giảm bớt tham số mô hình và tăng tốc thời gian huấn luyện.

3.3 Hàm Active Contour Loss đề xuất

Theo kết quả được trình bày ở [18], bằng cách tính gần đúng hàm đặc trưng từ một vectơ Heaviside với các ngưỡng multiphase [19], hàm Mumford-Shah có thể thu được như một hàm năng lượng phân biệt. Cụ thể hơn, đầu ra softmax kênh thứ n của mô hình DL được xây dựng như sau:

$$y_n(r) = \frac{e^{p_n(r)}}{\sum_{i=1}^{N} e^{p_n(r)}}, \qquad n = 1, 2, ..., N \tag{1}$$

trong đó $r \in \Omega \subset R^2$, $p_i(r)$ là output của mạng tại r từ layer trước khi softmax. Nhóm đã sử dụng hàm Mumford-Shah lấy ý tưởng từ CNN cho các tác vụ unsupervised:

$$L_{MScnn}(\theta; x) = \sum_{n=1}^{N} \int_{\Omega} \left| I(r) - c_n \right|^2 y_n(r) dr + \beta \sum_{n=1}^{N} \int_{\Omega} \left| \nabla y_n(r) \right| dr \tag{2}$$

trong đó:

$$c_n := c_n(\theta) = \frac{\int_{\Omega} x(r) y_n(r; \theta) dr}{\int_{\Omega} y_n(r; \theta) dr} \tag{3}$$

là giá trị pixel trung bình của lớp thứ n, $y_n(r) : = yn\ (r; \theta)$ là kết quả lớp softmax trong mô hình và $I(r)$ là phép đo hình ảnh đầu vào đã cho. Cuối cùng, θ biểu thị các tham số của mô hình có thể đào tạo.

Lấy ý tưởng từ [20], nhóm đề xuất hàm loss active contour cho multiphase segmentation:

$$L_{ac} = \frac{1}{P} \sum_{i=1}^{P} \sum_{n=1}^{N} \left(\left(d_{1n} - g_n(i) \right)^2 (1 - y_n(i)) + \left(d_{2n} - g_n(i) \right)^2 y_n(i) \right) \tag{4}$$

Trong đó P là số pixel của đầu vào, $g_n(i)$ biểu thị nhãn, d_{2n} là giá trị pixel trung bình của vùng thứ n và d_{1n} là pixel trung bình của tất cả vùng còn lại. Mỗi kênh của đầu ra đại diện cho một vùng được segmentation nhất định. Giả sử rằng pixel i không thuộc vùng thứ n, thuật ngữ $\left(d_{2n} - g_n(i) \right)^2 \approx 1$ và $\left(d_{1n} - g_n(i) \right)^2 \approx 0$, do đó,

mô hình cần giảm $y_n(i)$ để tối thiểu hóa L_{ac}. Tương tự, khi pixel i thuộc vùng thứ n, số hạng $\left(d_{2n} - g_n(i)\right)^2 \approx 0$ và $\left(d_{1n} - g_n(i)\right)^2 \approx 1$, do đó $y_n(i)$ được tăng lên. Dựa trên ý tưởng đã đề cập ở trên, chúng tôi cũng đề xuất hàm loss active contour cho multiphase segmentation như sau:

$$L_{uac} = -\frac{1}{P}\sum_{i=1}^{P}\sum_{n=1}^{N}\left(\left(d_{1n} - g_n(i)\right)^2 \log(y_n(i)+\varepsilon) + \left(d_{2n} - g_n(i)\right)^2 \log(1 - y_n(i)+\varepsilon)\right)$$

(5)

trong đó ε là tham số làm trơn để tránh logarithm explosion.

Logarithm này có thể phạt với trọng số lớn hơn khi mô hình dự đoán không chính xác với ground truth. Từ đó, chúng tôi chỉ ra hàm loss cho các tác vụ unsupervised:

$$L_{semi\_ac} = L_{ac} + \alpha L_{MScnn}(\theta; I) \quad (6)$$

$$L_{semi\_uac} = L_{uac} + \boldsymbol{\beta}\, L_{MScnn}(\theta; I) \quad (7)$$

Với α, β là các siêu tham số.

3.4 Kết quả và đánh giá

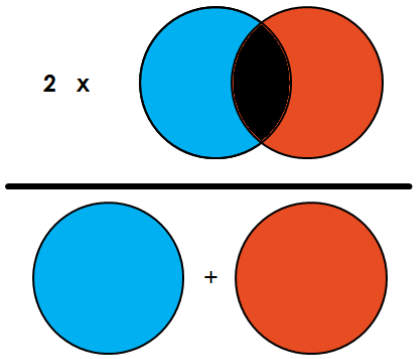
3.4.1 Độ đo đánh giá

Với bài toán segmentation, thì việc so khớp kích thước giữa ảnh phân vùng được từ mô hình và ground truth giúp đánh giá hiệu quả mô hình một cách trực quan nhất. Vì vậy nhóm đã sử dụng Dice Similarity Coefficient (DSC) và Jaccard Coefficient (JAC) để đánh giá hiệu suất của mạng nơ-ron sử dụng. Sau đây là công thức đánh giá:

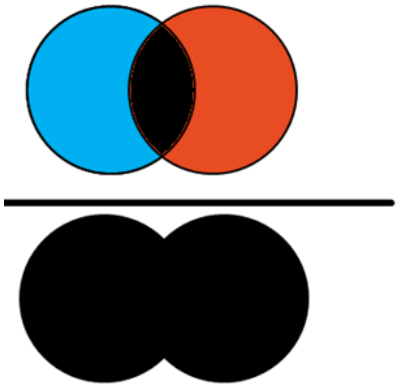
$$DSC = 2 \times \frac{TP}{FN + FP + 2 \times TP}$$

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

Với TP, FN, FP lần lượt là true positive, false negative, false positive so với ground truth. Do đó, đầu ra của mạng là ảnh binary segmentation để có thể dễ dàng so sánh với ground truth.



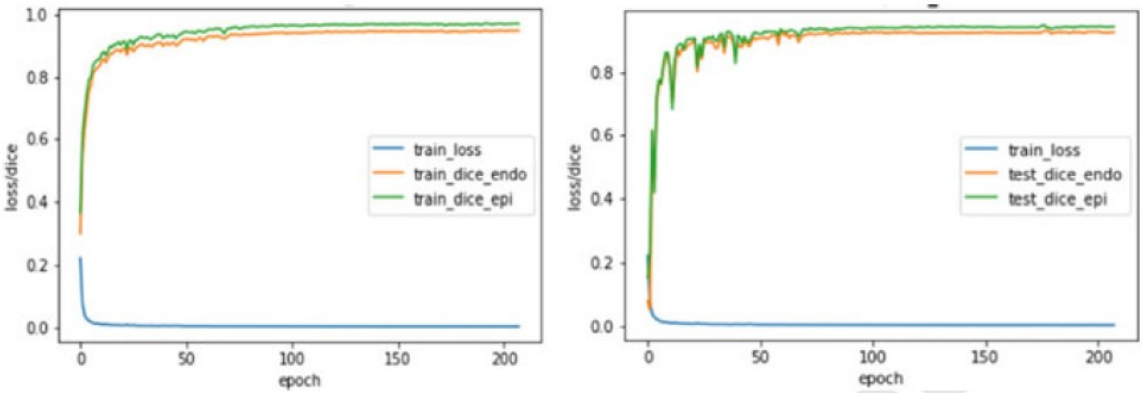
Hình 3.4 Minh họa DSC metric



Hình 3.5 Minh họa JAC(IoU) metric

3.4.2 Kết quả huấn luyện

Nhóm đã triển khai mô hình của mình và sử dụng thuật toán Nadam [24] để tối ưu hóa tham số huấn luyện của mô hình với learning rate ban đầu là 10^{-3}. Sau đó, quá trình huấn luyện được lặp lại trên bộ dữ liệu đào tạo ACDCA trong khoảng 300 epochs với batch size là 32. Learning rate cũng giảm 50% nếu validation score của không cải thiện sau khoảng 10 - 15 epochs (learning rate tối thiểu là 10^{-5}). Bên cạnh đó, nhóm đặt siêu tham số α, β lần lượt là 10^{-6}. và 1 trong tác vụ này. Sau đó, huấn luyện mạng nơ-ron của mình trên bốn giai đoạn chính với hàm loss như ở (4), (5), (6), (7).

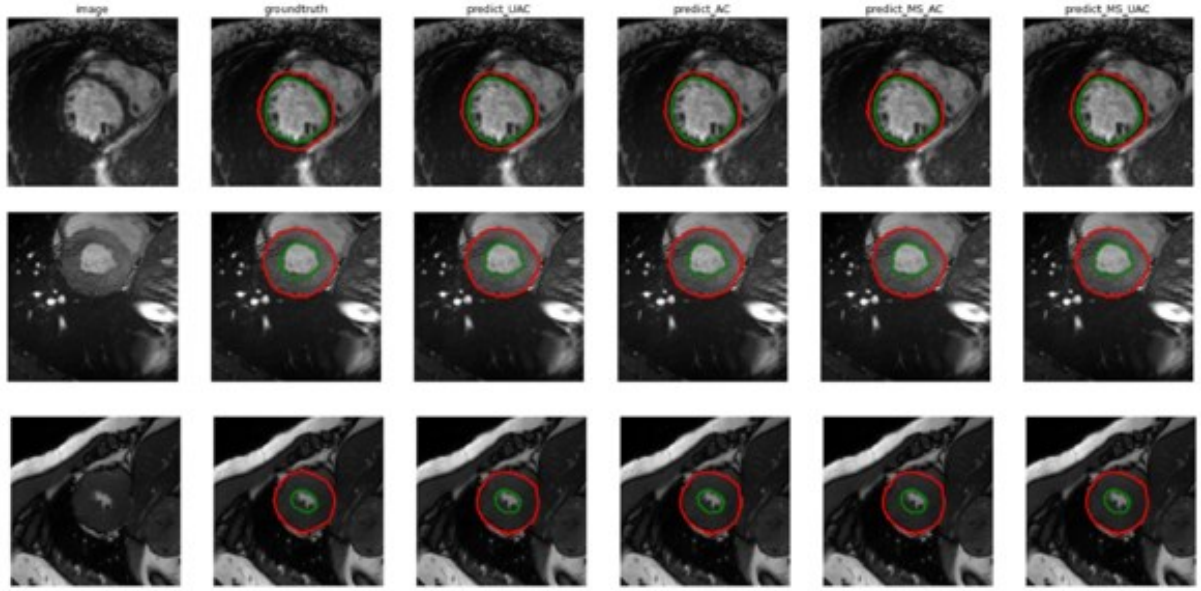


Hình 3.6 Learning curves của phương pháp đề xuất khi segmentation màng trong tim và màng ngoài tim với tập dữ liệu từ ACDCA

Learning curves cho thấy quá trình giảm của loss và hiệu suất segmentation trên tập train và bộ validation của mô hình của chúng tôi được hiển thị trong Hình 3.6. Ta có thể thấy rằng rằng tốc độ hội tụ mạng của bộ dữ liệu ACDCA rất nhanh (sau 200 epochs). DSC score của tập validation trong quá trình huấn luyện dao động lớn. Đó là bởi vì tập validation bao gồm một vài hình ảnh hoàn toàn khác với những hình ảnh

trong tập training. Do đó, trong lần học đầu tiên, mô hình này có một số khó khăn về segmentation những hình ảnh đó.

Hình 3.7 cung cấp cho chúng ta kết quả segmentation cơ bản cho một số hình ảnh trong bộ dữ liệu ACDCA sử dụng phương pháp đề xuất của chúng tôi. Có thể thấy trong biểu đồ này, các đường viền của màng trong tim và màng ngoài tim được segmentation khá tốt khi so sánh với ground truth.



Hình 3.7 Kết quả segmentation với đường màu xanh là màng trong tim và đường màu đỏ là màng ngoài tim

3.4.3 Đánh giá mô hình

Ngoài ra, để đánh giá định tính hiệu suất của phương pháp được đề xuất, chúng tôi đã trình bày Dice Similarity Coefficient và Jaccard Coefficient trung bình của phương pháp tiếp cận của chúng tôi và các phương pháp khác khi tất cả các hình ảnh được segmentation như trong Bảng 3.1. Từ bảng này, bằng so sánh định lượng, phương pháp đề xuất đạt được kết quả chính xác nhất.

Bảng 3.1 DSC và JAC của các phương pháp cơ bản và phương pháp với hàm loss đề xuất

Method	**Dice coefficient**		**Jacacd index**	
	Endo	**Epi**	**Endo**	**Epi**
FCN	0.89	0.92	0.83	0.89
SegNet	0.82	0.89	0.75	0.83
Unet	0.88	0.92	0.82	0.87
$\boldsymbol{L_{ac}}$	**0.92**	**0.93**	**0.88**	**0.90**
$\boldsymbol{L_{\text{uac}}}$	**0.93**	**0.93**	**0.89**	**0.90**
$\boldsymbol{L_{semi\_ac}}$	**0.93**	**0.94**	**0.89**	**0.91**
$\boldsymbol{L_{semi\_uac}}$	**0.93**	**0.95**	**0.89**	**0.92**

CHƯƠNG 4. KẾT LUẬN

Trong phương pháp này, nhóm đã đề xuất cải tiến hàm loss với active contour và biến thể của nó khi kết hợp với mạng nơ-ron sâu. Kết quả thử nghiệm trên bộ dữ liệu của ACDC cho thấy sự cải thiện đáng kể trong segmentation so với các lựa chọn thay thế hiện đại nhất. Đi xa hơn, đề xuất của chúng tôi về hàm loss này sẽ không chỉ được sử dụng trong phân đoạn hình ảnh y tế mà còn có thể được sử dụng trong các ứng dụng liên quan khác. Trong tương lai, chúng tôi dự định nghiên cứu tiềm năng của phương pháp của chúng tôi trong segmentation ảnh y tế nói riêng cũng như ảnh segmation nói chung.

TÀI LIỆU THAM KHẢO

1. B. Jähne, H. Haußecker, Computer vision and applications (2000)

2. T. Zhou, S. Ruan, S. Canu, A review: deep learning for medical image segmentation using multi-modality fusion. Array 3, 100004 (2019)

3. D. Mumford, J. Shah, Optimal approximations by piecewise smoothfunctions and associated variational problems. Commun. Pure Appl. Math. 42(5), 577–685 (1989)

4. T. Chan, L. Vese, Active contours without edges. IEEE Trans. Image Process. 10(2), 266–277 (2001)

5. A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. 40(1), 120–145 (2011)

6.A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation. IEEE J. Biomed. Health Inform. (2020)

7. Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39(6), 1856– 1867 (2019)

8. R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, Bi-directional ConvLSTM U-net with Densley connected convolutions, in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019

9. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer- Assisted Intervention* (Springer, 2015), pp. 234–241

10. S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3D fully convolutional deep networks, in *International Workshop on Machine Learning in Medical Imaging* (Springer, 2017), pp. 379–387

11. T.T. Tran, T.-T. Tran, Q.C. Ninh, M.D. Bui, V.-T. Pham, Segmentation of left ventricle in short-axis MR images based on fully convolutional network and active

contour model, in *International Conference on Green Technology and Sustainable Development* (Springer, 2020), pp. 49–59

12. S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19

13. D. Jha, M. Riegler, D. Johansen, P. Halvorsen, H. Johansen, Doubleu-net: a deep convolutional neural network for medical image segmentation, in *IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 558–564

14. X. Chen, B.M. Williams, S.R. Vallabhaneni, G. Czanner, R. Williams, Y. Zheng, Learning active contour models for medical image segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11632–11640

15. S.R. Hashemi, S.S.M. Salehi, D. Erdogmus, S.P. Prabhu, S.K. Warfield, A. Gholipour, Asym- metric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access 7, 1721–1735 (2018)

16. V.T. Pham, T.T. Tran, P.C. Wang, M.T. Lo, Tympanic membrane segmentation in otoscopic images based on fully convolutional network with active contour loss. Signal Image Video Process. https://doi.org/10.1007/s11760-020-01772-7 (2020)

17. S. Gur, L. Wolf, L. Golgher, P. Blinder, Unsupervised microvascular image segmentation using an active contours mimicking neural network, in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10722–10731

18. B. Kim, J.C. Ye, Mumford-Shah loss functional for image segmentation with deep learning. IEEE Trans. Image Process. 29, 1856–1866 (2019)

19. T.F. Chan, L.A. Vese, Image segmentation using level sets and the piecewise-constant Mumford-Shah model, in Tech. Rep. 0014, Computational Applied Math Group 2000. Citeseer

20. V.-T. Pham, T.-T. Tran, Active contour model and nonlinear shape priors with application to left ventricle segmentation in cardiac MR images. Optik 127(3), 991–1002 (2016)

21. M. Tan, Q.V. Le, Efficientnet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946 (2019)

22. B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network. arXiv:1505.00853 (2015)

23. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778

24. A. Tato, R. Nkambou, Improving adam optimizer (2018)

25. O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M.A.G. Ballester, Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging 37(11), 2514–2525 (2018)

26. P.V. Tran, A fully convolutional neural network for cardiac segmentation in short-axis MRI. arXiv:1604.00494 (2016)

27. V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495 (2017)