NGUYỄN MẠNH HIẾU

HỆ THỐNG THÔNG TIN

2020 – 2022

HÀ NỘI – NĂM 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN MẠNH HIẾU

PHÁT HIỆN VÀ NHẬN DẠNG HÌNH DÁNG LOẠI VIÊN THUỐC SỬ DỤNG DEEP LEARNING

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI - NĂM 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN MẠNH HIẾU

PHÁT HIỆN VÀ NHẬN DẠNG HÌNH DÁNG LOẠI VIÊN THUỐC SỬ DỤNG DEEP LEARNING

Chuyên ngành: HỆ THỐNG THÔNG TIN
Mã số: 8.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT
(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC : PGS. TSKH. HOÀNG ĐĂNG HẢI

HÀ NỘI - NĂM 2022

LỜI CAM ĐOAN

Tôi tên là Nguyễn Mạnh Hiếu, cam đoan: Luận văn Thạc sĩ Kỹ thuật "Phát hiện và nhận dạng hình dáng loại viên thuốc sử dụng Deep Learning" đây là công trình nghiên cứu của tác giả dưới sự hướng dẫn của PGS. TSKH. Hoàng Đăng Hải. Các kết quả nghiên cứu trong luận văn là trung thực, không sao chép bất kỳ từ một nguồn nào và dưới bất kỳ hình thức nào. Các nguồn tài liệu tham khảo đã được trích dẫn và ghi nguồn đúng quy định.

Tác giả của luận văn

Nguyễn Mạnh Hiếu

LỜI CẢM ƠN

Với lòng biết ơn sâu sắc, tôi xin gửi lời cảm ơn chân thành tới những người đã giúp đỡ tôi trong quá trình học tập, nghiên cứu khoa học.

Tôi xin chân thành cảm ơn:

Đầu tiên tôi xin cảm ơn thầy PGS. TSKH. Hoàng Đăng Hải đã tận tình hướng dẫn truyền đạt những kinh nghiệm quý báu và giúp đỡ tác giả từ những ngày bắt đầu hướng dẫn đến ngày bảo vệ.

Tiếp theo, tác giả cảm ơn các thầy cô trong Trường Học viện Công nghệ Bưu chính Viễn thông đã tận tình dạy dỗ, truyền đạt kiến thức quý báu.

Tôi xin trân trọng cảm ơn đơn vị nơi tôi công tác và làm việc đã tạo mọi điều kiện thuận lợi cho tôi trong suốt quá trình học cao học.

Cuối cùng, tác giả cảm ơn gia đình, đồng nghiệp, bạn bè đã luôn đồng hành, cổ vũ và giúp đỡ tác giả hoàn thành luận văn này.

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
RoI	Region of Interest	Vùng tập trung
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
ML	Machine Learning	Học máy
RPN	Region Proposal Network	Mạng đề xuất vùng
FCN	Fully connected network	Mạng kết nối đầy đủ
ANNs	Artificial neural networks	Mạng nơ-ron nhân tạo
ML	Machine Learning	Học máy

DANH SÁCH BẢNG

DANH MỤC HÌNH ẢNH

MỞ ĐẦU

1. Lý do chọn đề tài

Cùng với sự tiến bộ, phát triển của kinh tế - khoa học - xã hội, thế giới đang ngày càng tập trung hơn vào vấn đề chăm sóc sức khỏe cho con người; trong đó, việc nâng cao khả năng tự chăm sóc của người dân trở thành trọng tâm ở nhiều quốc gia. Việc tự chăm sóc sức khỏe tốt có thể giảm thiểu đáng kể các trường hợp phải nhập viện, phòng tránh các trường hợp có thể gây ra bệnh tật, nguy hiểm và trong một số trường hợp có thể dẫn tới tử vong. Tuy nhiên, với sự đa dạng của các loại thuốc viên trên thị trường, người bình thường thường, đặc biệt là những người lớn tuổi, thường gặp nhiều khó khăn trong việc phân biệt hoặc nhận dạng các viên thuốc không có nhãn mác, do đó tỉ lệ các trường hợp phải nhập viện hoặc gặp các vấn đề về sử dụng sai thuốc hiện nay tương đối cao. Theo [1], có 1/5 tác dụng phụ do thuốc gây ra có liên quan đến việc bệnh nhân sử dụng thuốc tại nhà; khoảng 18 triệu người từ 12 tuổi trở lên có thể sử dụng sai thuốc trị liệu tâm lý theo toa mỗi năm; 10% thuốc trên toàn cầu là hàng giả. Hậu quả sai sót về thuốc có thể ảnh hưởng nghiêm trọng đến sức khỏe và chất lượng cuộc sống của con người, dùng sai thuốc có thể dẫn đến các tác dụng phụ nghiêm trọng, nhập viện và thậm chí tử vong. Mỗi năm, tại Hoa Kỳ có khoảng hơn 7 triệu bệnh nhân bị ảnh hưởng, 7.000 đến 9.000 người chết do lỗi sử dụng thuốc [2], tại Ấn Độ có 5,2 triệu người chết do các sai sót y tế [3]… Chính vì vậy, việc phát triển các hệ thống phát hiện và nhận dạng hình dáng loại viên thuốc là một nhu cầu thực tiễn cấp thiết và đang là một chủ đề nghiên cứu được quan tâm nhiều (ví dụ [9-14]).

Mặc dù các viên thuốc có thể được phân biệt dựa trên đặc trưng cơ bản của các viên thuốc như: hình dạng, kích thước, màu sắc và dấu ấn riêng… nhưng vẫn cần những kiến thức chuyên môn nhất định và không tránh khỏi những nhầm lẫn chủ quan. Do đó, một hệ thống nhận dạng tự động có thể hỗ trợ hiệu quả cho cả bệnh nhân và nhân viên y tế trong việc phát hiện và nhận dạng các viên thuốc một cách chính xác, nhanh chóng và khách quan.

Đã có nhiều công trình nghiên cứu về nhận dạng và phân loại viên thuốc trong những năm qua, điển hình như [9-14]; trong đó, một số phương pháp phổ biến được áp dụng để phát hiện và phân loại viên thuốc là so khớp ảnh, hình dáng dựa vào trích

chọn các đặc trưng cơ bản của viên thuốc. Dựa trên nền tảng sử dụng công nghệ mạng nơ ron nhân tạo (Artificial Neuron Network - ANN) cho hiệu quả rất tốt đối với các bài toán nhận dạng đối tượng, các phương pháp học sâu (Deep Learning) với các mô hình mạng nơ ron tích chập (Convolutional Neural Network – CNN) đã được đề xuất nhằm tăng khả năng phát hiện, nhận dạng chính xác viên thuốc. Đã có nhiều mô hình CNN được đề xuất [4], điển hình là mô hình CNN kết hợp vùng (Region-CNN hay R-CNN), mô hình R-CNN nhanh (Fast R-CNN) [5], mô hình R-CNN nhanh hơn (Faster R-CNN) [6]. Một cải tiến mới của Faster R-CNN là Mask R-CNN được đề xuất năm 2017 [7]. Cho tới nay, Mask R-CNN là mô hình có nhiều ưu điểm vượt trội nhất trong số các mô hình R-CNN và cũng đã được đề xuất áp dụng cho mô hình nhận dạng viên thuốc, ví dụ [8]. Tuy nhiên, như đã chỉ ra trong [8], độ chính xác của Mask R-CNN phụ thuộc vào nhiều yếu tố như kỹ thuật phân đoạn ảnh, xác định vùng kết hợp, tập dữ liệu huấn luyện, môi trường kiểm thử.

Từ tình hình trên, việc nghiên cứu đề tài về ***"Phát hiện và nhận dạng hình dáng loại viên thuốc"*** là hết sức cần thiết, đáp ứng nhu cầu thực tiễn, góp phần xây dựng hệ thống phần mềm có khả năng hỗ trợ người dùng, nhân viên y tế, viện dưỡng lão nhằm nâng cao chất lượng chăm sóc sức khỏe người bệnh, phòng ngừa những rủi ro từ việc sử dụng sai loại thuốc.

2. Mục đích và nhiệm vụ nghiên cứu

- **Mục đích nghiên cứu:** Nghiên cứu, đề xuất giải pháp nhằm phát hiện và nhận dạng hình dáng loại viên thuốc một cách tự động bằng học sâu, thực nghiệm đánh giá giải pháp đã đề xuất.
- **Nhiệm vụ nghiên cứu:**
    - Xác định mục đích, yêu cầu phát hiện và nhận dạng hình dáng loại viên thuốc.
    - Nghiên cứu các phương pháp học sâu và khả năng áp dụng cho bài toán phát hiện và nhận dạng hình dáng loại viên thuốc.
    - Nghiên cứu xây dựng hệ thống (giải pháp) phát hiện và nhận dạng hình dáng loại viên thuốc bằng phương pháp học sâu.
    - Thực hiện thử nghiệm, đánh giá kết quả

3. Đối tượng và phạm vi nghiên cứu

*** Đối tượng:** Phương pháp phát hiện và nhận dạng hình dáng loại viên thuốc trong ảnh bằng học sâu.

*** Phạm vi nghiên cứu:**

- Các phương pháp phân đoạn ảnh và nhận dạng đối tượng.

- Phương pháp phân đoạn ảnh và nhận dạng hình dáng loại viên thuốc bằng học sâu.

- Xây dựng hệ thống phát hiện và nhận dạng hình dáng loại viên thuốc bằng học sâu.

- Thử nghiệm huấn luyện và đánh giá hệ thống trên bộ dữ liệu ảnh thuốc được lựa chọn và gán nhãn từ bộ dữ liệu ảnh CURE (chứa 8.973 hình ảnh của 196 lớp viên thuốc) [9].

4. Phương pháp nghiên cứu

* Phương pháp nghiên cứu lý thuyết: Khảo sát, nghiên cứu tài liệu, các bài báo, công trình nghiên cứu khoa học liên quan đã được công bố, qua đó tìm hiểu một số vấn đề chính, như:

- Lý thuyết liên quan vấn đề nghiên cứu.
- Tìm hiểu về một số phương pháp, mô hình học sâu hiện đại trong giải quyết bài toán nhận dạng đối tượng.
- Ứng dụng phát hiện và nhận dạng hình dáng loại viên thuốc.

* Phương pháp nghiên cứu thực nghiệm:

- Khảo sát, thu thập, xây dựng bộ dữ liệu ảnh thuốc mẫu.
- Lập trình ứng dụng thực nghiệm phương pháp đề xuất.
- Đánh giá kết quả thu được.

5. Kết quả đã đạt được của luận văn

- Phân tích yêu cầu bài toán nhận dạng hình dáng loại viên thuốc và khả năng áp dụng thực tiễn.
- Giải pháp áp dụng mạng nơ ron tích chập Mask R-CNN để tăng độ chính xác trong phát hiện và nhận dạng hình dáng loại viên thuốc.
- Kết quả thực nghiệm bài toán với Mask R-CNN và so sánh với một số phương pháp khác.

6. Cấu trúc của luận văn

Ngoài phần mở đầu, kết luận, danh mục tài liệu tham khảo và phụ lục, luận văn gồm 03 chương như sau:

- Chương 1 trình bày cơ sở lý thuyết về phát hiện và nhận dạng hình dáng loại viên thuốc.
- Chương 2 trình bày giải pháp hệ thống phát hiện và nhận dạng hình dáng loại viên thuốc sử dụng Mask R-CNN
- Chương 3 trình bày kết quả thực nghiệm và đánh giá.

Chương 1: CƠ SỞ LÝ THUYẾT

1.1. Bài toán phát hiện và nhận dạng loại viên thuốc

Phát hiện và nhận dạng loại viên thuốc đã là một nhu cầu thực tế từ nhiều năm qua và ngày càng gia tăng khi số lượng và chủng loại viên thuốc ngày càng lớn. Ngành công nghiệp dược phẩm đã và đang sản xuất ngày càng nhiều loại thuốc hiệu quả cho chữa bệnh. Các hãng sản xuất thuốc có xu hướng sản xuất thuốc theo cách riêng của họ. Tuy nhiên, do yêu cầu nhận dạng thuốc và chống thuốc giả mạo, các nhà sản xuất dược phẩm thường phải tuân theo những chuẩn chung về hình dạng, màu sắc, kích thước và dấu ấn của viên thuốc. Một nhu cầu thực tế đối với các nhà sản xuất dược phẩm là cần có hệ thống nhận dạng viên thuốc.

Mặt khác, việc sử dụng nhiều loại thuốc một cách thường xuyên trong đời sống ngày càng tăng ở mọi nhóm tuổi, đặc biệt là ở những người cao tuổi. Người dùng thường gặp khó khăn trong việc xác định thuốc khi thuốc được tách khỏi bao bì ban đầu và chuyển sang các hộp đựng khác nhau, hoặc chia nhỏ vào các hộp thuốc hàng ngày để sử dụng. Bên cạnh đó, các hiệu thuốc, dược sĩ, nhân viên y tế khi không xác định được đúng loại thuốc không rõ nguồn gốc cho bệnh nhân thì thường thông qua việc tìm kiếm trên không gian mạng một cách thủ công các đặc điểm vật lý (màu sắc, hình dạng và dấu ấn) hoặc nhận định dựa trên kiến thức, kinh nghiệm cá nhân. Đây là một quá trình chậm và dễ xảy ra sai sót, đòi hỏi sự khéo léo để xử lý những viên thuốc nhỏ, khả năng nhìn để đọc chữ viết nhỏ và một số mức độ hiểu biết chuyên môn nhất định. Những vấn đề trên đặt ra nhiều thách thức, rủi ro cho người dùng đối với việc sử dụng thuốc, nhất là trong môi trường bên ngoài bệnh viện [10]. Việc sử dụng sai loại viên thuốc đã và đang rất phổ biến. Theo thống kê trong [1], tại Mỹ từ năm 2003 đến năm 2007, số cuộc gọi đến các trung tâm kiểm soát chất độc đã tăng 44% với hầu hết trường hợp liên quan đến sử dụng sai thuốc viên. Một công cụ nhận dạng viên thuốc sẽ là công cụ hỗ trợ đắc lực cho những người cao tuổi, những bệnh nhân trong các cơ sở chăm sóc y tế.

Bài toán nhận dạng hình dáng loại viên thuốc là một bài toán con của nhận dạng đối tượng trong lĩnh vực xử lý ảnh, khoa học máy tính; các vấn đề ảnh hưởng

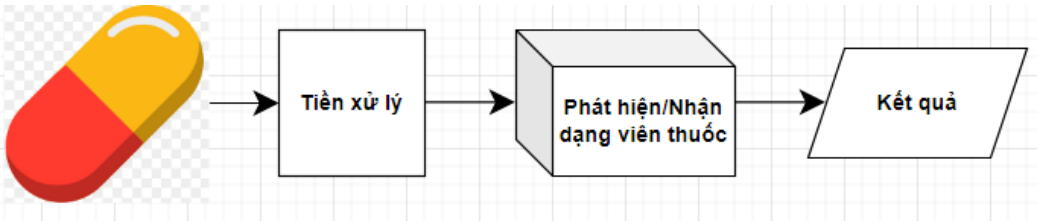
trực tiếp đến dữ liệu đầu vào trong điều kiện thực tế như: sự thay đổi trong điều kiện ánh sáng, việc sử dụng các ống kính máy ảnh khác nhau để chụp, nền chứa ảnh viên thuốc cùng màu với viên thuốc … có thể làm chênh lệch và biến đổi lớn trong các giá trị thông số ảnh thu được, đây là những thách thức cơ bản của bài toán này. Trước tình hình đó, nhiều hệ thống và công cụ nhận dạng viên thuốc đã được phát triển trong những năm qua. Chúng được phân loại chủ yếu thành hai loại chính là: hệ thống nhận dạng thủ công và hệ thống nhận dạng tự động [3].

* *Hệ thống nhận dạng thủ công:* yêu cầu người dùng cung cấp một cách thủ công các thông tin đầu vào về đặc điểm viên thuốc như: màu sắc, hình dạng, dấu ấn, v.v. Đã có một số trang web như WebMD, Pillbox, RxList, Drugs.com cho nhận dạng trực tuyến các viên thuốc. Tuy nhiên, các hệ thống này kém hiệu quả và còn có nhiều hạn chế về sai sót trong nhận dạng.

* *Hệ thống nhận dạng tự động:* không yêu cầu nhập dữ liệu theo cách thủ công, thực hiện nhận dạng viên thuốc tự động nhanh chóng và dễ sử dụng cho một số lượng lớn thuốc. Việc phát triển các hệ thống nhận dạng viên thuốc tự động đang trở thành một chủ đề nghiên cứu từ vài năm trở lại đây [9], [11], [12], [13], [14].

Trước nhu cầu thực tế về nhận dạng tự động viên thuốc, năm 2016 Thư viện Y khoa Hoa Kỳ (National Library of Medicine – NLM) đã tổ chức một cuộc thi với chủ đề "Thử thách nhận dạng hình ảnh viên thuốc" (Pill Image Recognition Challenge) [15]. Từ đó tới nay, nhiều giải pháp học sâu đã được đề xuất vào giải quyết bài toán nhận dạng thuốc viên, điển hình là các phương pháp được đề xuất trong [9-14], như: hệ thống nhận dạng thuốc viên bằng cách sử dụng GoogLeNet Inception Network với tính năng phát hiện cạnh Canny để cục bộ hóa thuốc viên của Wang và cộng sự [10]; mô hình mạng học sâu AlexNet của Wong và cộng sự [11]... Những nghiên cứu trên, nhất là thử thách của NLM đã thúc đẩy hơn nữa việc nghiên cứu giải quyết bài toán phát hiện và nhận dạng viên thuốc bằng những thành tựu, tiến bộ của khoa học công nghệ; cho đến nay bài toán đã ngày càng nhận được sự quan tâm của các nhà khoa học, công ty, doanh nghiệp và chính phủ cũng như người dân.

Bài toán phát hiện và nhận dạng loại viên thuốc bao gồm các phần chính như sau: Image → Tiền xử lý (Pre-processing) → Phát hiện/nhận dạng (Detection/Recognition) → Output Hình 1.1.



Hình 1.1. Thành phần chính của bài toán phát hiện, nhận dạng hình dáng viên thuốc

Trong đó, với dữ liệu đầu vào là ảnh chứa viên thuốc để trả về kết quả phát hiện, nhận dáng hình dáng loại viên thuốc, bài toán tập trung vào 2 bước chính như sau:

- Bước 1 - Tiền xử lý (Pre-processing): Đây là bước đầu tiên và hết sức quan trọng trong các bài toán học máy nói chung và bài toán phát hiện, nhận dáng hình dáng loại viên thuốc nói riêng, nhằm khoanh vùng, cắt ảnh, điều chỉnh kích thước, biến đổi (màu, độ sáng, lọc nhiễu, ,...) để giảm độ phức tạp, lọc nhiễu... qua đó trích xuất đặc trưng của hình ảnh giúp tạo dữ liệu đầu vào, hiệu quả nhất cho việc huấn luyện mô hình phát hiện và nhận dạng.
- Bước 2 - Phát hiện/nhận dạng (Detection/Recognition): Sau khi được tiền xử lý phù hợp với các kiến trúc của mô hình học máy, dữ liệu sẽ được sử dụng cho việc huấn luyện, xây dựng các bộ trọng số bởi các thuật toán học máy; khi đó mô hình xây dựng được sẽ có khả năng phát hiện, nhận dạng các đối tượng qua dữ liệu đã học để thể hiện ra cho người dùng.

1.2. Tiền xử lý dữ liệu

Phát hiện và nhận dạng hình dáng loại viên thuốc dựa trên việc xử lý dữ liệu đầu vào là các hình ảnh, do đó việc tiền xử lý dữ liệu - ở đây là tiền xử lý ảnh, trích xuất đặc trưng mặt không gian là quá trình tất yếu của mô hình định hướng xây dựng.

Tiền xử lý ảnh là một trong những phương pháp rất quan trọng giúp tăng cường dữ liệu cho huấn luyện. Một thuật toán phân loại hoặc phát hiện vật thể có thể được học đa dạng và tổng quát hơn nếu quá trình tiền xử lý dữ liệu tạo ra nhiều ảnh hơn cho nó. Tiền xử lý ảnh giúp tăng độ đa dạng của dữ liệu mẫu thông qua các kỹ thuật dịch chuyển ảnh về mặt hình học như: dịch chuyển phải, trái, lên, xuống, xoay ảnh, lật ảnh, biến đổi phối cảnh; các phương pháp lọc nhiễu, làm mờ thông qua bộ lọc; phương pháp phát hiện cạnh sử dụng bộ lọc Canny hoặc phương pháp ngưỡng; xác định các đường bao theo cường độ màu sắc tương đồng; xác định hộp giới hạn của vật thể; sử dụng thuật toán Non-Max Suppression để triệt tiêu các hộp giới hạn bị chồng lấn nhau...

Trên thực tế có nhiều trường hợp mô hình thể hiện trên dữ liệu huấn luyện cho kết quả rất tốt nhưng khi áp dụng vào thực tế lại không tốt. Nguyên nhân có thể đến từ khác biệt của hình ảnh huấn luyện và hình ảnh dự đoán về cường độ màu sắc ảnh, độ phân giải, chất lượng ảnh, mức độ chi tiết, mức độ bao quát,….

Do đó ứng dụng tiền xử lý dữ liệu sẽ giúp ta tăng độ đa dạng của dữ liệu mẫu, giải quyết được nhiều hơn các trường hợp trên thực tế và giúp nâng cao độ chính xác, cải thiện khả năng phát hiện, phân loại, nhận dạng viên thuốc. Cũng nhờ tiền xử lý ảnh mà tiết kiệm được chi phí thời gian cho quá trình chuẩn bị dữ liệu cho huấn luyện. Từ đó có thể tập trung vào việc tinh chỉnh cấu trúc và điều chỉnh mô hình để cải thiện chất lượng.

1.3. Phân đoạn ảnh và nhận dạng đối tượng

Phát hiện và nhận dạng hình dáng loại viên thuốc là bài toán cụ thể của lĩnh vực phân đoạn ảnh và nhận dạng đối tượng. Các công cụ phân loại hay phân lớp cũng như phân đoạn ảnh cung cấp một cách tiếp cận hiệu quả để trích xuất các đặc trưng từ hình ảnh, là công cụ, kỹ thuật không thể thiếu trong phát hiện và nhận dạng các đối tượng trong ảnh kỹ thuật số. Để nhận dạng đối tượng nói chung và viên thuốc nói riêng luôn cần áp dụng các kỹ thuật phát hiện đối tượng, phân đoạn ảnh hay phân loại ảnh, đây là bước thu thập những thông tin hữu ích nhất, các thông tin về hình dạng, đặc điểm quang phổ và không gian nhất định theo từng pixel và có thể được nhóm lại thành các đối tượng;

sau đó, các đối tượng có thể được nhóm lại thành các lớp đại diện cho các đối tượng trong thực tế.

1.3.1. Phân đoạn hình ảnh

Phân đoạn hình ảnh là một lĩnh vực cơ bản của thị giác máy tính được hỗ trợ bởi một lượng lớn nghiên cứu liên quan đến cả các thuật toán dựa trên xử lý hình ảnh và các kỹ thuật dựa trên học máy; qua đó góp phần xây dựng lên các phương pháp phát hiện đối tượng. Với sự phát triển của phân đoạn hình ảnh đã góp phần giải quyết một số thách thức trong phát hiện đối tượng, như:

- Trong phát hiện đối tượng, các hộp giới hạn luôn có hình chữ nhật. Vì vậy, chúng không giúp ích cho việc xác định hình dạng của đối tượng nếu đối tượng có chứa phần cong.
- Phát hiện đối tượng không thể ước tính chính xác một số phép đo như diện tích của đối tượng, chu vi của đối tượng từ hình ảnh.

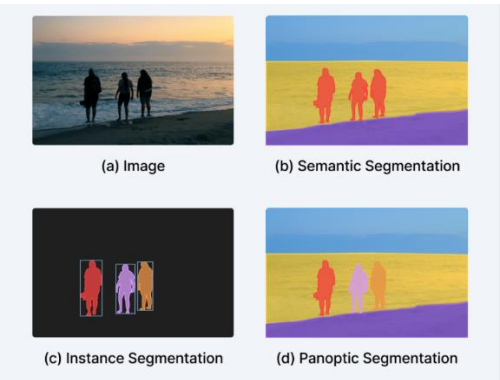
Cùng với việc trở thành một trong những lĩnh vực quan trọng nhất của thị giác máy tính, phân đoạn hình ảnh cũng là một trong những vấn đề lâu đời nhất nhận được nhiều sự quan tâm của các nhà nghiên cứu, với các công trình đầu tiên liên quan đến các kỹ thuật phát triển vùng (region based) nguyên thủy và các phương pháp tối ưu hóa được phát triển từ những năm 1970 – 1972 [18].

Phân đoạn hình ảnh là một miền phụ của thị giác máy tính và xử lý hình ảnh kỹ thuật số nhằm mục đích nhóm các vùng hoặc phân đoạn tương tự của hình ảnh dưới nhãn lớp tương ứng của chúng. Vì toàn bộ quá trình là kỹ thuật số nên việc tạo ra một biểu diễn của hình ảnh tương tự dưới dạng pixel làm cho nhiệm vụ hình thành các phân đoạn tương đương với nhiệm vụ nhóm các pixel. Phân đoạn hình ảnh là một phần mở rộng của ***phân loại hình ảnh***, ngoài việc phân loại, các nghiên cứu thực hiện xác định vị trí đối tượng. Do đó, phân đoạn hình ảnh là một tập hợp siêu phân loại hình ảnh với mô hình xác định chính xác nơi có đối tượng tương ứng bằng cách vạch ra ranh giới của đối tượng.

Giống như tất cả các thuật toán học sâu giám sát, quy trình phân đoạn được giám sát yêu cầu **dữ liệu được chú thích** quy mô lớn để huấn luyện. Loại chú thích được yêu cầu thay đổi tùy theo loại phân đoạn được thực hiện bởi mô hình, từ các chú thích rất cụ thể được yêu cầu trong các nhiệm vụ phân đoạn theo từng đối tượng (Instance Segmentation) đến các chú thích rất đơn giản được yêu cầu trong các tác vụ phân đoạn ngữ nghĩa (Semantic Segmentation). Loại chú thích được yêu cầu và độ chính xác cần thiết khác nhau tùy theo các trường hợp sử dụng mô hình và bản đồ phân đoạn. Tập dữ liệu chú thích cho các tác vụ như phân đoạn ngữ nghĩa rất dễ xây dựng trong khi chú thích cho phân đoạn toàn cảnh khó hơn vì chúng yêu cầu xem xét sự chồng chéo giữa các đối tượng. Ví dụ, các trường hợp sử dụng như hình ảnh y tế và xe tự hành yêu cầu chú thích chính xác cao hơn để phân đoạn so với các ứng dụng đơn giản khác.

1.3.2. Các cách tiếp cận trong phân đoạn hình ảnh

Các nhiệm vụ phân đoạn hình ảnh có thể được phân thành ba nhóm dựa trên số lượng và loại thông tin mà chúng truyền tải [19]. Trong khi phân đoạn ngữ nghĩa phân đoạn ra một ranh giới rộng lớn của các đối tượng thuộc một lớp cụ thể thì phân đoạn theo từng đối tượng cung cấp một bản đồ phân đoạn cho mỗi đối tượng mà nó xem xét trong hình ảnh mà không có bất kỳ ý tưởng nào về lớp mà đối tượng đó thuộc về; phân đoạn khái quát (Panoptic segmentation) cho đến nay là cung cấp nhiều thông tin nhất, là sự kết hợp của các nhiệm vụ phân đoạn theo ngữ nghĩa và theo từng đối tượng. Phân đoạn khái quát cung cấp bản đồ phân đoạn của tất cả các đối tượng của bất kỳ lớp cụ thể nào hiện diện trong hình ảnh. Cụ thể:



Hình 1.2. Các loại phân đoạn hình ảnh [19]

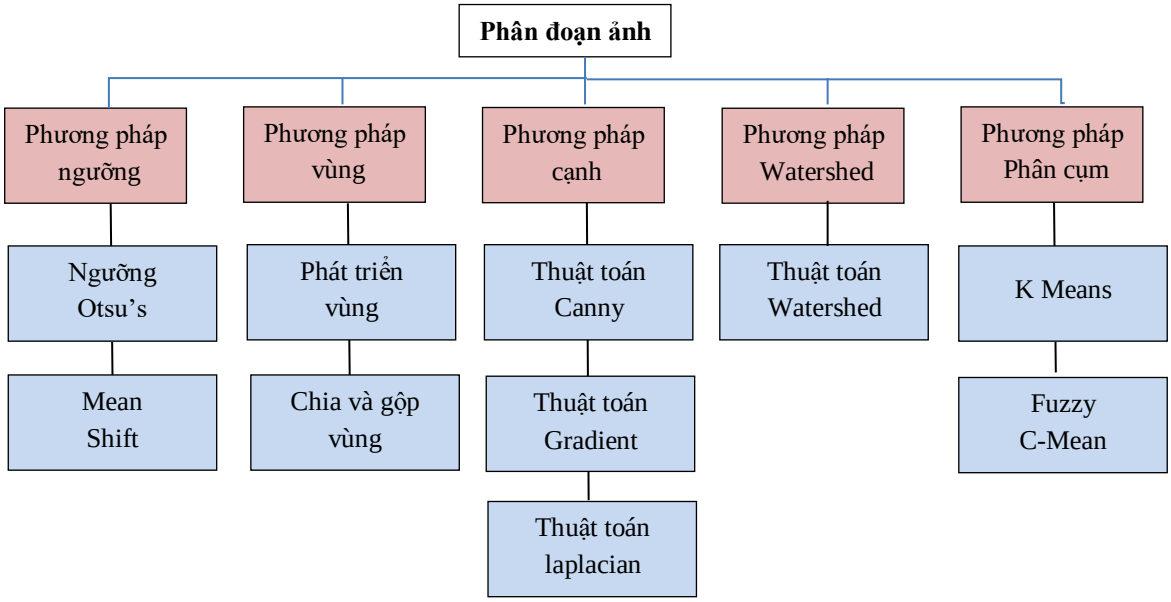
- **Phân đoạn ngữ nghĩa - Phân đoạn theo loại đối tượng (Semantic Segmentation):** Phân đoạn ngữ nghĩa đề cập đến việc phân loại các pixel trong một hình ảnh thành các lớp ngữ nghĩa. Các điểm ảnh thuộc một lớp cụ thể được phân loại đơn giản vào lớp đó mà không tính đến thông tin hoặc bối cảnh khác. Như có thể mong đợi, đây là một câu lệnh vấn đề được xác định kém khi có nhiều trường hợp được nhóm chặt chẽ của cùng một lớp trong hình ảnh. Hình ảnh một đám đông trên đường phố sẽ có mô hình phân đoạn ngữ nghĩa dự đoán toàn bộ khu vực đám đông thuộc lớp "người đi bộ", do đó cung cấp rất ít chi tiết hoặc thông tin chuyên sâu về hình ảnh.
- **Phân đoạn phiên bản - Phân đoạn từng đối tượng (Instance Segmentation):** Các mô hình phân đoạn thực thể phân loại pixel thành các loại trên cơ sở "cá thể" chứ không phải là các lớp. Một thuật toán phân đoạn cá thể không có ý tưởng về lớp mà một vùng đã phân loại thuộc về nhưng có thể tách các vùng đối tượng chồng chéo hoặc rất giống nhau trên cơ sở ranh giới của chúng. Nếu hình ảnh tương tự của một đám đông mà chúng ta đã nói trước đây được đưa vào một mô hình phân đoạn cá thể, thì mô hình sẽ có thể tách riêng từng người khỏi đám đông cũng như các đối tượng xung quanh (lý tưởng là), nhưng sẽ không thể dự đoán từng vùng / đối tượng là một ví dụ của.
- **Phân đoạn khái quát:** Hay phân đoạn theo sơ đồ là nghiên cứu phân đoạn được phát triển gần đây nhất, có thể được biểu thị bằng sự kết hợp giữa phân đoạn ngữ nghĩa và phân đoạn đối tượng trong đó mỗi thể hiện của một đối tượng trong hình ảnh được tách biệt và danh tính của đối tượng được dự đoán. Các thuật toán phân đoạn Panoptic cho thấy khả năng ứng dụng quy mô lớn trong các nhiệm vụ phổ biến như ô tô tự lái, nơi phải thu thập một lượng lớn thông tin về môi trường xung quanh ngay lập tức với sự trợ giúp của luồng hình ảnh.

1.4. Các kỹ thuật phân đoạn ảnh

Phân đoạn hình ảnh bắt đầu từ xử lý hình ảnh kỹ thuật số kết hợp với các thuật toán tối ưu hóa, đáng chú ý như Hình 1.3 [20].

1.4.1. Kỹ thuật phân đoạn theo ngưỡng

Ngưỡng là một trong những phương pháp phân đoạn hình ảnh dễ dàng nhất trong đó ngưỡng được đặt để chia pixel thành hai lớp. Các pixel có giá trị lớn hơn giá trị ngưỡng được đặt thành 1 trong khi các pixel có giá trị nhỏ hơn giá trị ngưỡng được đặt thành 0. Do đó, hình ảnh được chuyển đổi thành một bản đồ nhị phân, dẫn đến quá trình này thường được gọi là mã hóa nhị phân. Ngưỡng hình ảnh rất hữu ích trong trường hợp sự khác biệt về giá trị pixel giữa hai lớp mục tiêu là rất cao và có thể dễ dàng chọn một giá trị trung bình làm ngưỡng. Ngưỡng thường được sử dụng để mã hóa hình ảnh để các thuật toán tiếp theo như phát hiện và nhận dạng đường viền chỉ hoạt động trên hình ảnh nhị phân có thể được sử dụng.



Hình 1.3. Một số phương pháp phân đoạn ảnh truyền thống [20]

1.4.2. Kỹ thuật phân đoạn theo vùng

Các thuật toán phân đoạn dựa trên vùng hoạt động bằng cách tìm kiếm những điểm tương đồng giữa các pixel liền kề và nhóm chúng thành một lớp chung. Thông thường, quy trình phân đoạn bắt đầu với một số pixel được đặt làm pixel gốc và thuật toán hoạt động bằng cách phát hiện ranh giới ngay lập tức của các pixel gốc và phân loại chúng là tương tự hoặc khác nhau. Những pixel lân cận ngay sau đó được coi như hạt giống và các bước được lặp lại cho đến khi toàn bộ hình ảnh được phân

đoạn. Một ví dụ về thuật toán tương tự là thuật toán ***watershed*** phổ biến để phân đoạn ảnh bằng cách bắt đầu từ cực đại cục bộ của bản đồ khoảng cách Euclide và phát triển theo ràng buộc rằng không có hai hạt giống nào có thể được phân loại là thuộc cùng một vùng hoặc bản đồ phân đoạn.

1.4.3. Kỹ thuật phân đoạn theo cạnh

Phân đoạn cạnh [21], còn được gọi là phát hiện cạnh, là nhiệm vụ phát hiện các cạnh trong ảnh. Từ quan điểm dựa trên phân đoạn, chúng ta có thể nói rằng phát hiện cạnh tương ứng với việc phân loại pixel nào trong hình ảnh là pixel cạnh và tách các pixel cạnh đó theo một lớp riêng biệt một cách tương ứng. Phát hiện cạnh thường được thực hiện bằng cách sử dụng các bộ lọc đặc biệt cung cấp cho chúng ta các cạnh của hình ảnh khi tích chập. Các bộ lọc này được tính toán bởi các thuật toán chuyên dụng hoạt động trên việc ước tính độ dốc hình ảnh theo tọa độ x và y của mặt phẳng không gian. Dưới đây là một ví dụ về phát hiện cạnh bằng thuật toán phát hiện cạnh Canny, một trong những thuật toán phát hiện cạnh phổ biến nhất.

Hình 1.4. Phân đoạn cạnh [21]

1.4.4. Kỹ thuật phân đoạn theo phân cụm

Các thủ tục phân đoạn hiện đại phụ thuộc vào các kỹ thuật xử lý ảnh thường sử dụng các thuật toán phân cụm [20] để phân đoạn. Các thuật toán phân cụm hoạt động tốt hơn so với các thuật toán của chúng và có thể cung cấp các phân đoạn hợp lý tốt trong một khoảng thời gian nhỏ. Các thuật toán phổ biến như thuật toán phân cụm K-mean là thuật toán không được giám sát hoạt động bằng cách nhóm các pixel có các thuộc tính chung lại với nhau như thuộc về một phân đoạn cụ thể. Đặc biệt,

phân cụm K-mean là xem xét tất cả các pixel và phân cụm chúng thành “k” lớp. Khác với các phương pháp phân đoạn theo vùng, các phương pháp dựa trên phân nhóm không cần điểm giống để bắt đầu phân đoạn.

1.4.5. Kỹ thuật phân đoạn dựa trên học sâu

Các mô hình phân đoạn theo ngữ nghĩa cung cấp các bản đồ phân đoạn dưới dạng kết quả đầu ra tương ứng với các đầu vào mà chúng được cung cấp. Các bản đồ phân đoạn này thường được phân thành *n* kênh với *n* là số lớp mà mô hình phải phân đoạn. Mỗi kênh trong số *n* kênh này có bản chất là nhị phân với các vị trí đối tượng được "lấp đầy" bằng các kênh và các vùng trống bao gồm các số 0. Bản đồ cơ bản là một mảng số nguyên kênh đơn có cùng kích thước với đầu vào và có phạm vi "n", với mỗi phân đoạn được "lấp đầy" bằng giá trị chỉ số của các lớp tương ứng (các lớp được lập chỉ mục từ 0 đến n-1). Đầu ra mô hình ở định dạng nhị phân “n - kênh” còn được gọi là biểu diễn được mã hóa *one-hot* hai chiều của các dự đoán. Mạng nơ-ron thực hiện phân đoạn thường sử dụng cấu trúc bộ mã hóa-giải mã (encoder-decoder); trong đó, bộ mã hóa được theo sau bởi một nút cổ chai (bottleneck) và một bộ giải mã hoặc các lớp lấy mẫu trực tiếp từ nút cổ chai (như trong FCN).

U-Net, **DeepLab** của Facebook đóng vai trò là một cột mốc quan trọng, cung cấp các kết quả hiện đại về phân đoạn ngữ nghĩa. DeepLab đã sử dụng các cụm phức tạp thay thế cho các hoạt động gộp đơn giản và ngăn ngừa mất mát thông tin đáng kể trong khi downsampling. Họ đã giới thiệu thêm về tính năng trích xuất đặc trưng đa tỷ lệ với sự trợ giúp của Atrous Spatial Pyramid Pooling để giúp mạng phân đoạn các đối tượng bất kể kích thước của chúng. Để khôi phục thông tin ranh giới, một trong những phần quan trọng nhất của phân đoạn ngữ nghĩa cũng như phiên bản, họ đã sử dụng các Trường ngẫu nhiên có điều kiện (CRF) được kết nối đầy đủ. Kết hợp độ chính xác bản địa hóa chi tiết của CRF, khả năng nhận dạng của CNN đã giúp DeepLab cung cấp bản đồ phân đoạn có độ chính xác cao, đánh bại các phương pháp như FCN và SegNet một cách rõ ràng. Các báo cáo công trình nghiên cứu phân đoạn ảnh như SegNet, U-Net và DeepLab đã đặt nền móng cho các công trình trong tương lai như **Mask-R-CNN** và các hoạt động như PspNet và GSCNN.

Phân đoạn hình ảnh là một bước quan trọng trong thị giác nhân tạo. Máy móc cần chia dữ liệu trực quan thành các phân đoạn để quá trình xử lý phân đoạn cụ thể diễn ra. Do đó, phân đoạn hình ảnh được tìm thấy trong các lĩnh vực nổi bật như robot, hình ảnh y tế, xe tự hành và phân tích video thông minh… Ngoài các ứng dụng này, phân đoạn hình ảnh cũng được sử dụng bởi vệ tinh trên hình ảnh trên không để phân đoạn các con đường, tòa nhà và cây cối.

1.5. Phân loại đối tượng

Phân loại đối tượng trong hình ảnh là một trong những nhiệm vụ cơ bản nhất trong thị giác máy tính. Nó đã tạo ra một cuộc cách mạng và thúc đẩy những tiến bộ công nghệ trong các lĩnh vực nổi bật nhất, bao gồm công nghiệp ô tô, chăm sóc sức khỏe, sản xuất, v.v. Phân loại đối tượng là nhiệm vụ liên kết một (phân loại nhãn đơn) hoặc nhiều nhãn (phân loại đa nhãn) với các đối tượng trong ảnh nhất định [22].

*** Phân loại nhãn đơn:** Phân loại nhãn đơn là nhiệm vụ phân loại phổ biến nhất trong phân loại ảnh có giám sát. Như tên cho thấy, sẽ có một nhãn hoặc chú thích có sẵn cho mỗi hình ảnh trong phân loại nhãn đơn. Do đó, mô hình xuất ra một giá trị hoặc dự đoán duy nhất cho mỗi hình ảnh mà nó xử lý. Đầu ra từ mô hình là một vectơ có độ dài bằng số lớp và giá trị biểu thị điểm số của hình ảnh thuộc lớp này.

Một hàm kích hoạt (activate function) Softmax thường được sử dụng để đảm bảo đầu ra dự đoán (phân lớp) của mô hình nằm trong đoạn [0, 1], được tính theo công thức (1.1).

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_i)} \tag{1.1}$$

Một số ví dụ về bộ dữ liệu phân loại nhãn đơn bao gồm MNIST, SVHN, ImageNet, v.v. Phân loại nhãn đơn có thể thuộc loại phân loại đa phân lớp (trong đó có nhiều hơn 2 lớp) hoặc phân loại nhị phân (trong đó số lượng lớp bị hạn chế chỉ có 2).

*** Phân loại đa nhãn:** Phân loại đa nhãn là một nhiệm vụ phân loại trong đó mỗi hình ảnh có thể chứa nhiều hơn một nhãn và một số hình ảnh có thể đồng thời chứa tất cả các nhãn. Mặc dù điều này có vẻ giống với phân loại nhãn đơn về một

khía cạnh nào đó, nhưng đây là vấn đề phức tạp hơn so với phân loại nhãn đơn. Nhiệm vụ phân loại đa nhãn tồn tại phổ biến trong lĩnh vực hình ảnh y tế, khi một bệnh nhân có thể mắc nhiều bệnh cần được chẩn đoán từ dữ liệu hình ảnh dưới dạng tia X; hay một bức ảnh thuốc do người dung cung cấp có thể có nhiều loại thuốc. Hơn nữa, trong môi trường tự nhiên, việc đánh nhãn hình ảnh cũng có thể được nhận định như một bài toán phân loại nhiều nhãn, giúp chỉ ra các đối tượng hiện diện trong hình ảnh.

Mô hình phân loại đa nhãn xuất ra một vectơ có độ dài bằng số lớp mẫu, với các giá trị là độ chính xác của từng lớp được tình bởi các hàm kích hoạt khác nhau, như: Softmax (mô hình càng gần với 1 thì đó là lớp chính xác); hay khi dữ liệu đầu ra độc lập với nhau vì nhiều lớp có thể tồn tại cùng một lúc thì không thể áp dụng Softmax vì sẽ tạo ra các vấn đề làm giảm độ chính xác của mô hình, như sự thiên vị (bias) cho giá trị lớn nhất; trong trường hợp này có thể áp dụng hàm Sigmoid, Relu…

1.6. Mạng nơ ron nhân tạo

1.6.1. Khái quát

Mạng nơ-ron nhân tạo - Artificial neural networks (ANNs) là một mô hình học máy (ML), hệ thống tính toán lấy cảm hứng từ mạng nơ-ron sinh học cấu thành não động vật [22]. ANN dựa trên một tập hợp các đơn vị hoặc nút được kết nối được gọi là tế bào thần kinh nhân tạo, mô hình hóa lỏng lẻo các tế bào thần kinh trong não sinh học. Mỗi kết nối, giống như khớp thần kinh trong não sinh học, có thể truyền tín hiệu đến các tế bào thần kinh khác. Một tế bào thần kinh nhân tạo nhận một tín hiệu sau đó xử lý và có thể phát tín hiệu cho các tế bào thần kinh kết nối với chúng. "Tín hiệu" tại một kết nối là một số thực và đầu ra của mỗi nơ-ron được tính bằng một số hàm phi tuyến tính của tổng các đầu vào của nó. Các kết nối được gọi là các cạnh; các tế bào thần kinh và các cạnh thường có trọng số sẽ điều chỉnh khi quá trình huấn luyện diễn ra. Trọng số làm tăng hoặc giảm cường độ của tín hiệu tại một kết nối. Tế bào thần kinh có thể có ngưỡng sao cho tín hiệu chỉ được gửi đi khi tín hiệu tổng hợp vượt qua ngưỡng đó. Thông thường, các tế bào thần kinh được tập hợp thành các lớp. Các lớp khác nhau có thể thực hiện các phép biến đổi khác nhau trên các đầu vào của

chúng. Tín hiệu đi từ lớp đầu tiên (lớp đầu vào), đến lớp cuối cùng (lớp đầu ra), có thể sau khi đi qua các lớp nhiều lần.

Việc huấn luyện ANN: Mạng nơ-ron nhân tạo được huấn luyện bằng cách xử lý các ví dụ, mỗi ví dụ chứa một "đầu vào" và "kết quả" đã biết, tạo thành các liên kết có trọng số xác suất giữa hai mạng, được lưu trữ trong cấu trúc dữ liệu của chính mạng đó. Việc huấn luyện mạng nơ-ron từ một ví dụ nhất định thường được tiến hành bằng cách xác định sự khác biệt giữa đầu ra đã xử lý của mạng (thường là một dự đoán) và đầu ra mục tiêu; sự khác biệt này là lỗi. Sau đó, mạng sẽ điều chỉnh các liên kết có trọng số của nó theo một quy tắc huấn luyện và sử dụng giá trị lỗi này. Các điều chỉnh liên tiếp sẽ khiến mạng nơ-ron tạo ra đầu ra ngày càng giống với đầu ra mục tiêu. Sau khi có đủ số lượng điều chỉnh này, việc huấn luyện có thể được chấm dứt dựa trên các tiêu chí nhất định. Các hệ thống như vậy "huấn luyện" để thực hiện các nhiệm vụ bằng cách xem xét các ví dụ mà không được lập trình với các quy tắc dành riêng cho nhiệm vụ.

1.6.1. Mạng nơ-ron tích chập

Các phương pháp học máy truyền thống (chẳng hạn như máy nhận thức đa lớp, máy vectơ hỗ trợ, v.v.) hầu hết sử dụng “cấu trúc nông” để xử lý một số mẫu và đơn vị tính toán hạn chế. Khi các đối tượng mục tiêu phong phú, hiệu suất và khả năng khái quát hóa của các bài toán phân loại phức tạp rõ ràng là không đủ. Là một phân hệ của kiến trúc mạng nơ-ron nhân tạo, mạng nơ-ron tích chập (Convolutional Neural Network - CNN) được phát triển trong những năm gần đây đã được sử dụng rộng rãi trong lĩnh vực xử lý ảnh vì hiệu quả trong xử lý tốt các vấn đề phân loại, nhận dạng ảnh và đã mang lại sự cải thiện lớn về độ chính xác của nhiều tác vụ học máy. CNN đã trở thành một mô hình học sâu phổ biến và mạnh mẽ; lần đầu tiên được phát triển và sử dụng vào khoảng những năm 1980 [4]. Điều tối đa mà CNN có thể làm vào thời điểm đó là nhận dạng các chữ số viết tay; chủ yếu được sử dụng trong các lĩnh vực bưu chính để đọc mã zip, mã pin, v.v. Điều quan trọng cần nhớ về bất kỳ mô hình học sâu nào là nó đòi hỏi một lượng lớn dữ liệu để huấn luyện và cũng đòi hỏi nhiều tài nguyên máy tính. Đây là một nhược điểm lớn đối với CNN vào thời

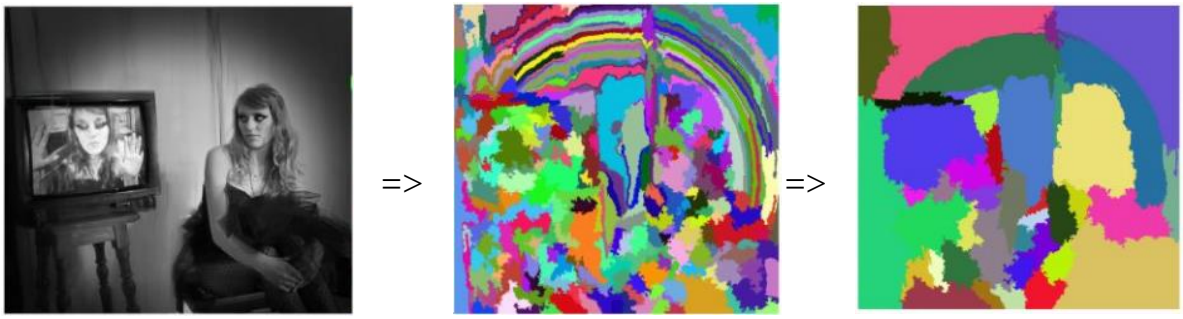
kỳ đó và do đó CNN chỉ giới hạn trong lĩnh vực bưu chính và không thể bước vào thế giới máy học. Năm 2012, Alex Krizhevsky nhận ra rằng đã đến lúc phải đưa ngành học sâu trở lại sử dụng mạng nơ-ron nhiều lớp. Sự sẵn có của các bộ dữ liệu lớn, cụ thể hơn là các bộ dữ liệu ImageNet với hàng triệu hình ảnh được gắn nhãn và nguồn tài nguyên máy tính dồi dào đã cho phép các nhà nghiên cứu hồi sinh CNN. Trong học sâu, mạng nơ-ron tích chập (CNN/ConvNet) là một lớp mạng nơ-ron sâu, được áp dụng phổ biến nhất để phân tích hình ảnh trực quan.

Một mạng nơ-ron tích chập bao gồm một lớp đầu vào, các lớp ẩn và một lớp đầu ra. Trong mọi mạng nơron lan truyền thẳng (Feed-forward Neural Network - FNN), các lớp giữa được gọi là lớp ẩn vì các đầu vào và đầu ra của chúng bị che bởi hàm kích hoạt và tích chập cuối cùng, các lớp ẩn bao gồm các lớp thực hiện các phép chập. Sản phẩm đầu ra là sản phẩm bên trong Frobenius và hàm kích hoạt của nó có thể là ReLU, Sigmod... Khi hạt nhân (kernel) tích chập trượt dọc theo ma trận đầu vào, phép toán tích chập tạo ra một bản đồ đặc trưng, bản đồ này sẽ đóng góp vào đầu vào của lớp tiếp theo. Tiếp theo là các lớp khác như lớp gộp, lớp được kết nối đầy đủ và lớp chuẩn hóa. Hiện nay có nhiều kiến trúc khác nhau của CNN, là chìa khóa trong việc xây dựng các thuật toán mạnh mẽ, như: LeNet, AlexNet, VGGNet, GoogLeNet, ResNet, ZFNet… Xuất phát từ CNN, các nhà nghiên cứu đã cho ra đời các mô hình mạng tích chập hiện đại hiệu quả hơn trong phát hiện đối tượng, như: R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN.

1.6.2. Mô hình R-CNN

Mô hình R-CNN đề xuất một loạt các hộp trong hình ảnh và kiểm tra xem có bất kỳ hộp nào trong số này chứa bất kỳ đối tượng nào hay không. R-CNN sử dụng tìm kiếm có chọn lọc để trích xuất các hộp từ một hình ảnh (các hộp này được gọi là vùng). Về cơ bản, có bốn vùng hình thành một đối tượng: tỷ lệ, màu sắc, kết cấu và vùng bao quanh khác nhau. Tìm kiếm có chọn lọc xác định các mẫu này trong hình ảnh và dựa trên đó, đề xuất các vùng khác nhau. Đầu tiên, mô hình R-CNN lấy một hình ảnh làm đầu vào, sau đó, tạo các phân đoạn con ban đầu để chúng ta có nhiều vùng từ hình ảnh này từ đó kết hợp các vùng tương tự để tạo thành một vùng lớn hơn

(dựa trên sự tương đồng về màu sắc, sự tương đồng về kết cấu, sự tương đồng về kích thước và sự tương thích về hình dạng); cuối cùng, các vùng này tạo ra các vị trí đối tượng cuối cùng (vùng quan tâm), Hình 1.5.



Hình 1.5. Ví dụ ứng dụng mạng R-CNN

** Vấn đề với R-CNN:* Cho đến nay, R-CNN hữu ích để phát hiện đối tượng nhưng kỹ thuật này đi kèm với những hạn chế riêng của nó. Việc huấn luyện một mô hình R-CNN rất tốn kém và chậm do:

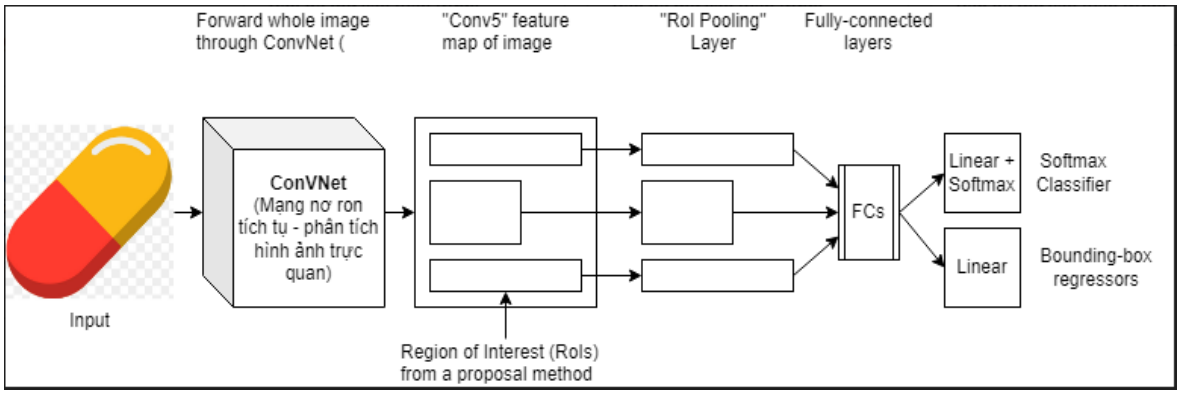
- Cần trích xuất 2.000 vùng cho mỗi hình ảnh dựa trên tìm kiếm có chọn lọc.
- Trích xuất các đặc trưng sử dụng CNN cho mọi vùng hình ảnh. Giả sử chúng ta có N hình ảnh, thì số lượng đặc trưng của CNN sẽ là N*2.000
- Toàn bộ quá trình phát hiện đối tượng bằng R-CNN có ba mô hình:
    - CNN để trích xuất đặc trưng
    - Bộ phân loại SVM tuyến tính để xác định các đối tượng
    - Mô hình hồi quy để xác định các hộp giới hạn.

Tất cả các quá trình này kết hợp với nhau làm cho R-CNN rất chậm. Mất khoảng 40-50 giây để đưa ra dự đoán cho mỗi hình ảnh mới, điều này về cơ bản làm cho mô hình trở nên cồng kềnh và thực tế không thể xây dựng khi đối mặt với một tập dữ liệu khổng lồ.

1.6.3. Mô hình Fast R-CNN

Thay vì thực hiện mô hình CNN 2.000 lần cho mỗi hình ảnh, chúng ta có thể thực hiện CNN chỉ một lần cho mỗi hình ảnh và nhận được tất cả các vùng quan tâm (vùng chứa một số đối tượng). Ross Girshick, tác giả của R-CNN, đã đưa ra ý tưởng chạy CNN chỉ một lần cho mỗi hình ảnh và sau đó tìm cách chia sẻ tính toán đó trên

2.000 khu vực [5]. Trong Fast R-CNN, cung cấp hình ảnh đầu vào cho CNN, từ đó tạo ra các bản đồ đối tượng cục bộ phức hợp. Sử dụng các bản đồ này, các khu vực đề xuất (RoI) được trích xuất. Sau đó, sử dụng lớp tổng hợp RoI để định hình lại tất cả các vùng được đề xuất thành một kích thước cố định, để nó có thể được đưa vào một mạng được kết nối đầy đủ. Vì vậy, thay vì sử dụng ba mô hình khác nhau (như trong R-CNN), Fast R-CNN sử dụng một mô hình duy nhất trích xuất các đặc trưng từ các vùng, chia chúng thành các lớp khác nhau và trả về các hộp ranh giới cho các lớp được xác định đồng thời, Hình 1.6.

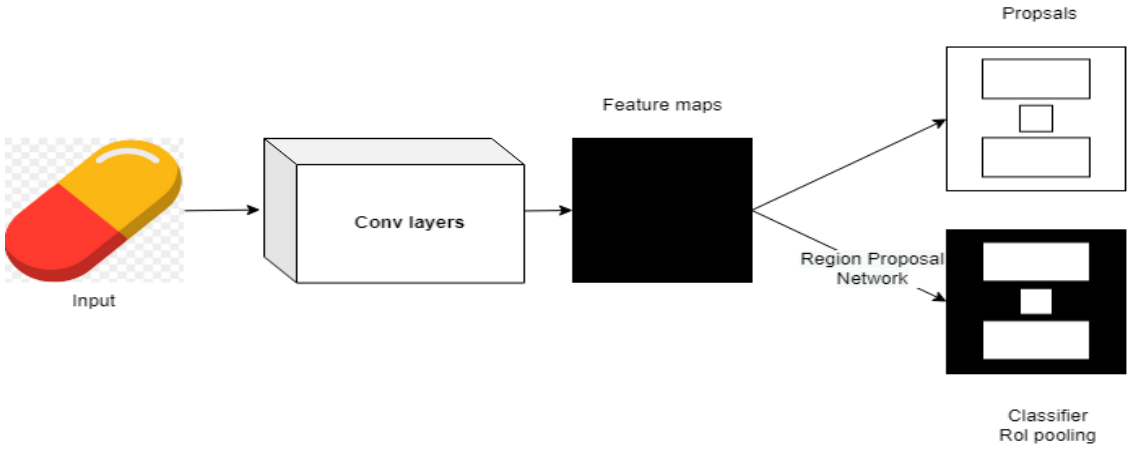


Hình 1.6. Kiến trúc tổng quan của mô hình Fast R-CNN

* *Các vấn đề với Fast R-CNN:* Mô hình này cũng sử dụng tìm kiếm có chọn lọc như một phương pháp đề xuất để tìm vùng quan tâm (RoI), đây là một quá trình chậm và tốn thời gian. Mất khoảng 2 giây cho mỗi hình ảnh để phát hiện đối tượng, tốt hơn nhiều so với R-CNN nhưng khi chúng ta xem xét các bộ dữ liệu lớn trong cuộc sống thực, thì Fast R-CNN vẫn chưa đáp ứng được.

1.6.4. Mô hình Faster R-CNN

Faster R-CNN là phiên bản sửa đổi của Fast R-CNN [6]. Sự khác biệt chính giữa chúng là Fast R-CNN sử dụng tìm kiếm có chọn lọc để tạo vùng quan tâm, trong khi Faster R-CNN sử dụng "Mạng đề xuất khu vực" (Region Proposal Network - RPN) lấy bản đồ đặc trưng hình ảnh làm đầu vào và tạo một tập hợp các đối tượng đề xuất, mỗi đề xuất có một điểm đối tượng làm đầu ra, Hình 1.7.



Hình 1.7. Mô hình luồng Faster R-CNN

Cụ thể, để bắt đầu, Faster R-CNN lấy bản đồ đối tượng cục bộ từ CNN và chuyển đến Mạng đề xuất khu vực RPN, sử dụng cửa sổ trượt trên các bản đồ đặc trưng này và tại mỗi cửa sổ, tạo ra ***k**-anchor-box* có hình dạng và kích thước khác nhau. *Anchor box* là các hộp ranh giới có kích thước cố định được đặt trong toàn bộ hình ảnh và có các hình dạng và kích thước khác nhau. Đối với mỗi anchor box, RPN dự đoán hai điều:

- Đầu tiên là xác suất mà một anchor box là một đối tượng (Faster R-CNN không xem xét đối tượng đó thuộc về lớp nào);
- Thứ hai là bộ hồi quy hộp giới hạn để điều chỉnh các anchor phù hợp hơn với đối tượng.

Kết quả là các hộp giới hạn có hình dạng và kích thước khác nhau được chuyển đến lớp tổng hợp RoI nhằm lấy từng đề xuất và cắt nó để mỗi đề xuất chứa một đối tượng; nó trích xuất các bản đồ đặc trưng có kích thước cố định cho mỗi anchor box. Sau đó, các bản đồ đặc trưng này được chuyển đến một lớp fully connection có hàm softmax và một lớp hồi quy tuyến tính. Cuối cùng nó phân loại và dự đoán các hộp giới hạn cho các đối tượng.

* *Vấn đề của Faster R-CNN:* Tất cả các thuật toán phát hiện đối tượng từ CNN đến Faster R-CNN đều sử dụng các vùng để xác định các đối tượng. Mạng không xem toàn bộ hình ảnh trong một lần mà tập trung vào các phần của hình ảnh một cách tuần tự. Điều này tạo ra hai vấn đề:

- Thuật toán yêu cầu nhiều lần đi qua một hình ảnh duy nhất để trích xuất tất cả các đối tượng.
- Vì có các hệ thống khác nhau lần lượt hoạt động, hiệu suất của các hệ thống phía sau phụ thuộc vào cách các hệ thống trước đó hoạt động

1.6.5. Mạng Mask R-CNN

1.6.5.1. Sự ra đời của Mask R-CNN

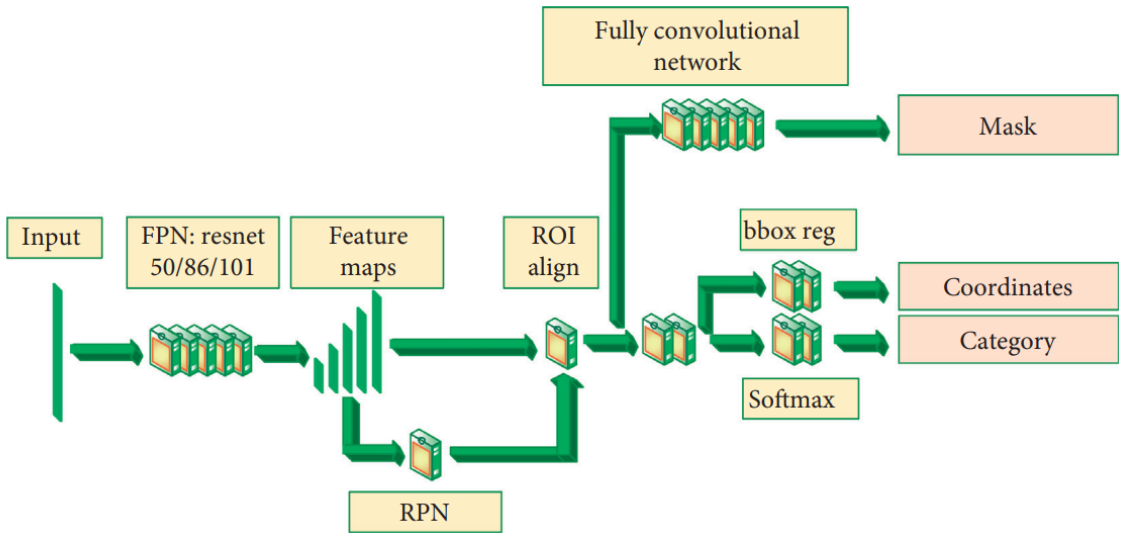
Hai phương pháp chính để phát hiện đối tượng là phương pháp tiếp cận dựa trên máy học và phương pháp tiếp cận dựa trên học sâu. Khi các yêu cầu về độ chính xác và tốc độ của thuật toán tiếp tục tăng lên, các thuật toán nhận dạng đối tượng như R-CNN, Fast R-CNN, Faster R-CNN và Mask R-CNN đã được đề xuất. Mask R-CNN được đề xuất vào năm 2017, không bổ sung bất kỳ kỹ năng nào nhưng Mask R-CNN vượt trội hơn tất cả các mô hình nhận dạng đơn lẻ vào thời điểm đó và đánh bại nhà vô địch năm 2016 trong thử thách trên tập dữ liệu Microsoft COCO. Theo yêu cầu thực tế, Mask R-CNN có ưu điểm là tốc độ nhanh và độ chính xác cao trong các nhiệm vụ phát hiện đối tượng, nó được ứng dụng vào nhiều lĩnh vực. Bảng 1.1 so sánh những đặc trưng cơ bản của họ mô hình CNN [23].

Bảng 1.1. So sánh R-CNN, Fast R-CNN, Faster R-CNN và Mask R-CNN

<table>
<tr><th>Nền tảng mạng</th><th>R-CNN</th><th>Fast R-CNN</th><th>Faster R-CNN</th><th>Mask R-CNN</th></tr>
<tr><td>Thời gian đề xuất</td><td>2014</td><td>2015</td><td>2016</td><td>2017</td></tr>
<tr><td>Đề xuất vùng</td><td>Selective search</td><td>Selective search</td><td>RPN</td><td>RPN</td></tr>
<tr><td>Trích chọn đặc trưng</td><td>CNN</td><td rowspan="2">CNN + ROI pool</td><td rowspan="2">CNN + ROI pool</td><td rowspan="2">CNN + ROI align</td></tr>
<tr><td>Phân lớp đặc trưng</td><td>SVM</td></tr>
<tr><td>Chức năng</td><td>Phân lớp, phát hiện</td><td>Phân lớp, phát hiện</td><td>Phân lớp, phát hiện</td><td>Phân lớp, phát hiện, phân đoạn</td></tr>
</table>

Thời gian xử lý trên 1 ảnh	47 seconds	2 seconds	0.2 seconds	0.2 seconds
mAP (VOC 2012)	62.4%	68.4%	70.4%	—

1.6.5.2. Kiến trúc mô hình và đặc trưng của Mask R-CNN



Hình 1.8. Mô hình luồng của Mask R-CNN [7]

Mục tiêu của luận văn này là phát triển một hệ thống hỗ trợ cho phân đoạn cá thể tới từng viên thuốc. Vì vậy khắc phục những thách thức của phân đoạn đối tượng yêu cầu phát hiện chính xác tất cả các đối tượng trong một hình ảnh đồng thời phân đoạn chính xác từng đối tượng, hệ thống cần kết hợp các yếu tố từ các nhiệm vụ thị giác máy tính cổ điển của phát hiện đối tượng với mục tiêu là phân loại mỗi pixel thành một nhóm danh mục cố định.

Mask R-CNN là sự mở rộng của Faster R-CNN bằng cách thêm một nhánh để dự đoán mặt nạ (mask) phân đoạn trên từng Khu vực quan tâm (RoI) song song với nhánh hiện có để phân loại và hồi quy hộp giới hạn [7], **Error! Reference source not found.**.

Mask R-CNN rất đơn giản để thực hiện và huấn luyện dựa trên khung Faster R-CNN, tạo điều kiện cho một loạt các thiết kế kiến trúc linh hoạt:

- Nhánh mặt nạ là một lớp kết nối toàn bộ (Fully Connected Network – FCN) nhỏ được áp dụng cho mỗi RoI, dự đoán mặt nạ phân đoạn theo cách pixel-to-pixel. Ngoài ra, nhánh mặt nạ chỉ thêm một chi phí tính toán nhỏ, cho phép xây dựng một hệ thống nhận dạng tốc độ cao.

- Quan trọng nhất, Faster R-CNN không được thiết kế để căn chỉnh pixel-to-pixel giữa đầu vào và đầu ra của mạng; điều này thể hiện rõ nhất trong cách RoI Pool là hoạt động cốt lõi để tham gia vào các ConvNet, thực hiện lượng tử hóa dữ liệu thô để khai thác đặc trưng. Để khắc phục sự sai lệch, Mask R-CNN đề xuất một lớp đơn giản, không có lượng tử hóa, được gọi là RoIAlign, bảo toàn chính xác các vị trí các pixel. RoIAlign có tác động lớn trong Mask R-CNN, nó cải thiện độ chính xác của mặt nạ tương đối từ 10% đến 50%. Đồng thời, Mask R-CNN tách dự đoán mặt nạ và lớp, tiến hành dự đoán mặt nạ nhị phân cho từng lớp một cách độc lập, không có sự đan xen, ảnh hưởng giữa các lớp và dựa vào nhánh phân loại RoI của mạng để dự đoán danh mục. Ngược lại, FCN thường thực hiện phân loại nhiều lớp theo pixel, kết hợp giữa phân đoạn và phân loại khiến hoạt động kém hiệu quả đối với phân đoạn.

1.6.5.3. Hoạt động của Mask R-CNN

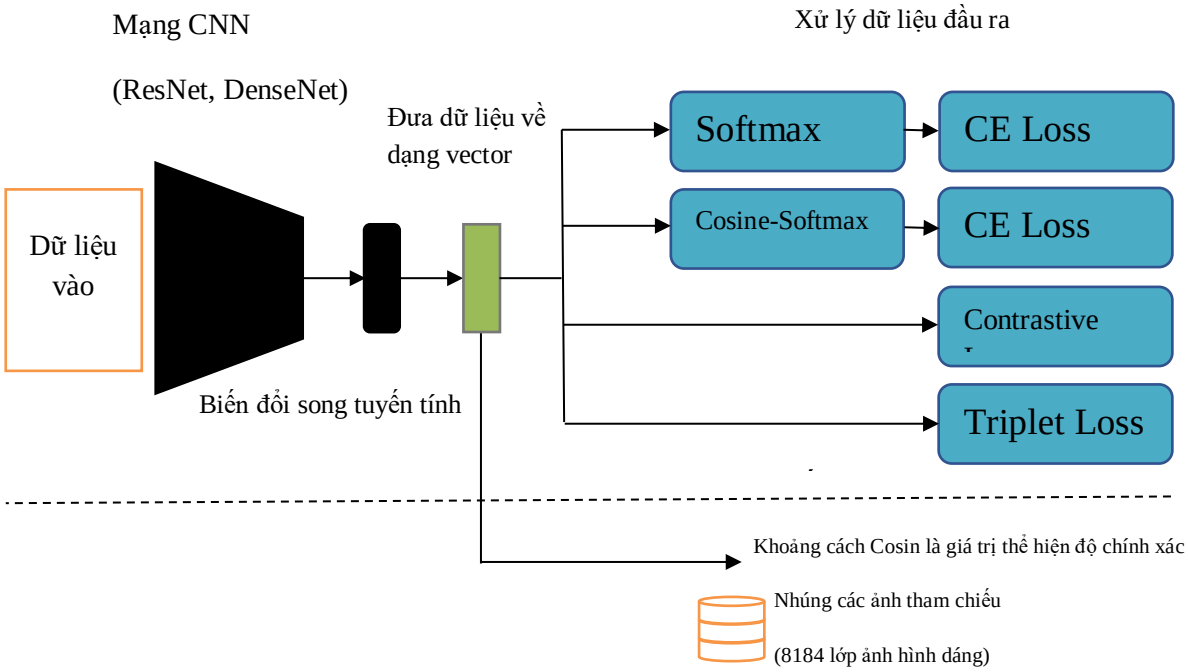
Mask R-CNN áp dụng quy trình hai giai đoạn giống nhau khi đều sử dụng các mạng RPN trong giai đoạn đầu; trong giai đoạn thứ hai, tiến hành song song với việc dự đoán độ lệch lớp và hộp, Mask R-CNN cũng xuất ra một mặt nạ nhị phân cho mỗi RoI. Điều này trái ngược với hầu hết các hệ thống họ CNN khác, khi việc phân loại đối tượng phụ thuộc vào kết quả dự đoán về mặt nạ. Trong quá trình đào tạo, việc xác định giá trị *loss* trên mỗi RoI được lấy mẫu là $L = L_{cls} + L_{box} + L_{mask}$. Trong đó, L_{cls} và L_{box} như được định nghĩa trong kiến trúc mạng Faster R-CNN; đồng thời, nhánh mặt nạ có một $K\ m^2$ - đầu ra cho mỗi RoI được áp dụng hàm *sigmoid* trên mỗi pixel và xác định L_{mask} là giá trị *loss* entropy chéo nhị phân trung bình.

Mask R-CNN dự đoán một mặt nạ kích thước $m \times m$ từ mỗi RoI bằng cách sử dụng các mạng kết nối hoàn toàn nhỏ FCN. Điều này cho phép mỗi lớp trong nhánh mặt nạ duy trì $m \times m$ bố cục không gian đối tượng mà không thu gọn nó thành một biểu diễn vectơ thiếu kích thước không gian.

1.7. Một số nghiên cứu liên quan

Nhiều nghiên cứu đang được thực hiện trong lĩnh vực nhận dạng viên thuốc tự động; trong đó, nổi bật một số phương pháp như sau:

(1) Tháng 9/2020, Naoto Usuyama và cộng sự [24] dùng bộ dữ liệu gồm 13.000 hình ảnh viên thuốc (kích thước 224x224 điểm ảnh, tăng cường dữ liệu: xoay và thay đổi góc nhìn) thuộc 9.804 lớp (hai mặt "front - back" cho 4.902 viên thuốc khác nhau, trong đó, ảnh thực tế người dùng chụp là 960 loại) từ bộ dữ liệu NIH để xây dựng mô hình học chuyển giao từ Resnet152 và DenseNet kết hợp sử dụng B-CNN và BCP ở bước cuối để phân loại viên thuốc. Hình 1.9 mô tả tổng quan mô hình của phương pháp.

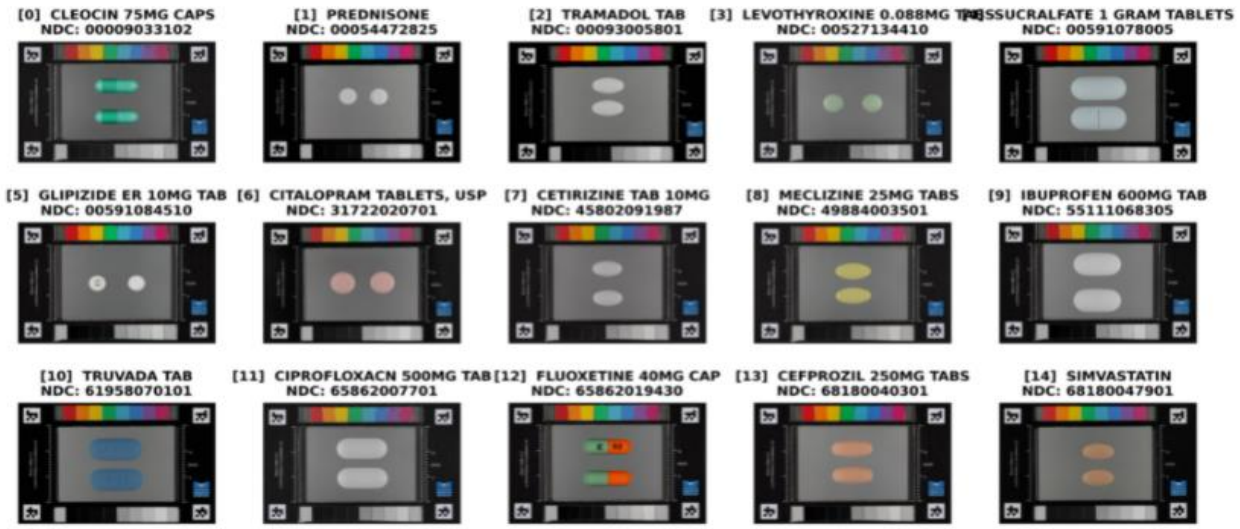


Hình 1.9. Mô hình phương pháp kết hợp ResNet, DenseNet và B-CNN/BCP

Kết quả thực nghiệm, các tác giả thu được hiệu quả đạt 85% mAP, 82% gAP trên dữ liệu là ảnh các viên thuốc có cả hai mặt. *Ưu điểm,* mô hình thu được có độ chính xác tương đối cao, là một trong những phương pháp hiệu quả hiện nay trong giải quyết bài toán nhận dạng viên thuốc; đã sử dụng các mô hình học sâu hiện đại (Resnet152, DenseNet) nâng cao hiệu quả của mô hình. *Tuy nhiên, phương pháp tồn tại một số nhược điểm như:* sử dụng bộ dữ liệu NIH mặc dù có số lượng lớn hình ảnh

viên thuốc, nhưng số lượng ảnh thuốc cho mỗi lớp là rất ít (thường là 2 ảnh mặt trước và sau cho 01 lớp), do đó hiệu quả nhận dạng chưa tối ưu, chưa áp dụng hiệu quả đối với dữ liệu thực tế người dùng cung cấp; đồng thời, chưa tiến hành thử nghiệm trên trường hợp nhiều viên thuốc cùng xuất hiện và có sự chồng chéo, đè lấp nhau, do đó chưa đánh giá được hết trường hợp thực tế.

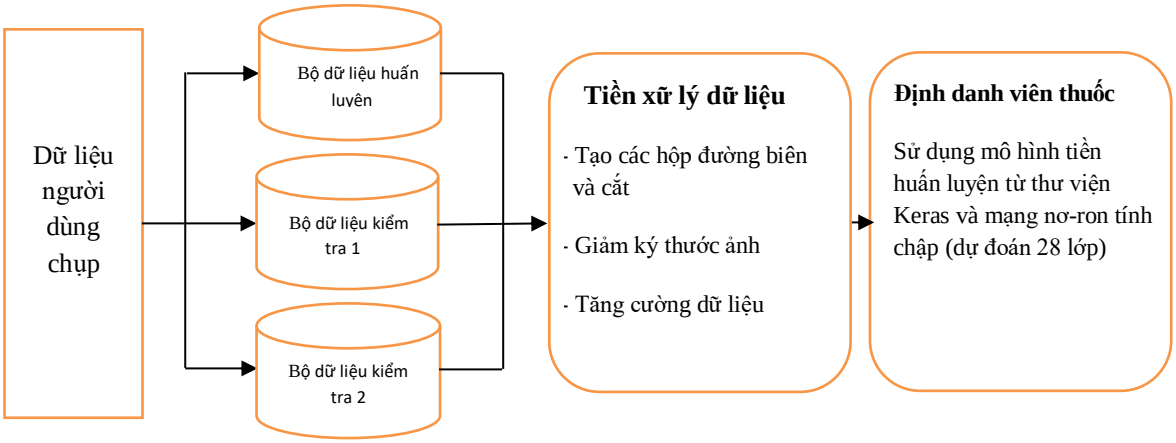
(2) Tháng 5/2020, Alphonso Woodbury [11] xây dựng bộ dữ liệu gồm 15 lớp hình ảnh viên thuốc với 490 hình ảnh từ bộ dữ liệu NIH để tiến hành thực nghiệm thông qua phương pháp học chuyển giao từ mô hình VGG-16 nhận dạng viên thuốc, Hình 1.10 minh họa bộ dữ liệu của tác giả xây dựng.



Hình 1.10. Bộ dữ liệu 15 lớp, 490 hình ảnh thuốc của Alphonso Woodbury

Kết quả, tác giả thu được độ chính xác lên đến **93%** trên bộ dự liệu kiêm tra và sấp xỉ **50%** trên bộ dữ liệu ảnh thuốc với nền thực do người dùng chụp. *Ưu điểm,* tác giả đã thu thập, phân loại dữ liệu từ bộ NIH để tăng số lượng ảnh viên thuốc mỗi lớp, nâng cao độ chính xác trên tập dữ liệu kiểm tra (đạt 93%). *Tuy nhiên, tồn tại một số nhược điểm như:* mô hình được học chuyển giao từ mô hình VGG-16, đây chưa phải là mô hình hiệu quả nhất trong phát hiện và nhận dạng đối tượng hiện nay, do đó cần nâng cấp, ứng dụng từ các mô hình học sâu mới; nghiên cứu này cũng chưa kiểm thử trên các ảnh có sự chồng chéo, đè nấp của nhiều viên thuốc, do đó chưa sát với thực tế sử dụng của người dùng.

(3) Năm 2020, Suwat Tangwattananuwat và cộng sự [13] đã sử dụng bộ dữ liệu gồm 3,074 hình ảnh của 28 loại thuốc áp dụng học chuyển giao từ mô hình VGG16, Inception-Resnet-V2 và Xception. Hình 1.11 dưới đây mô tả về mô hình hệ thống thực nghiệm của nhóm tác giả.

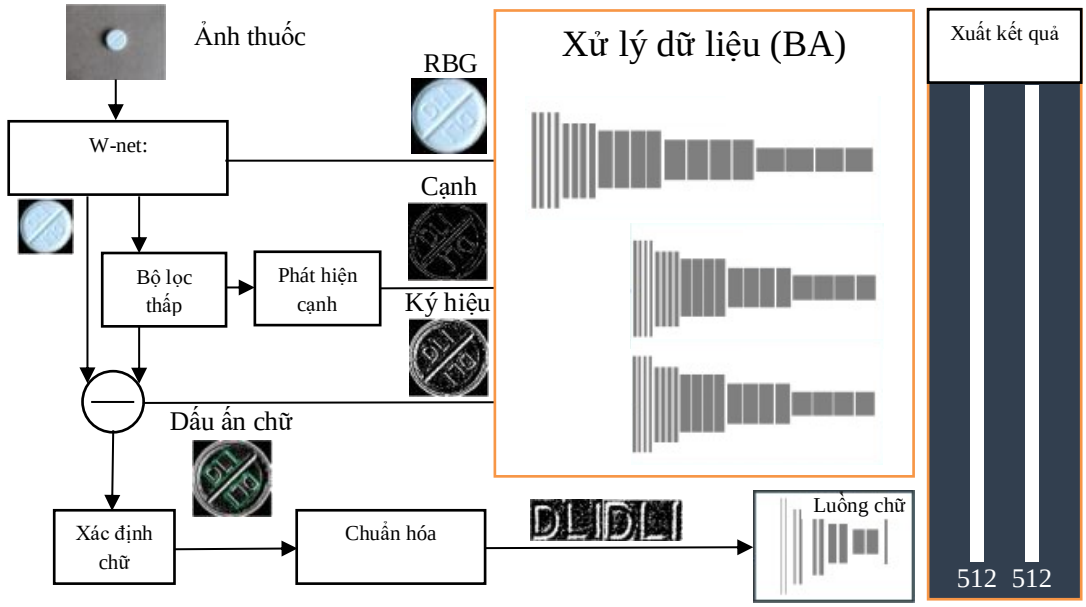


Hình 1.11. Mô hình hệ thống thực nghiệm định danh viên thuốc của Suwat

Kết quả thực nghiệm, cả 3 mô hình VGG16, Inception-Resnet-V2 và Xception đều cho độ chính xác 100% của tập dữ liệu kiểm tra 1 và lần lượt là 71,42%, 82,14% và 77,38% của tập dữ liệu kiểm tra 2. *Ưu điểm*, tác giả đã thu thập bộ dữ liệu lớn các ảnh viên thuốc, ứng dụng các mô hình học sâu tương đối hiệu quả trong phát hiện và nhận dạng đối tượng; kết quả thu được trên bộ dữ liệu kiểm tra tương đối cao (trên 70%). *Tuy nhiên tồn tại một số nhược điểm như*: Do sử dụng bộ dữ liệu không công khai nên khó so sánh với các phương pháp khác; kết quả thu được chưa phải là kết quả tốt nhất so với nhiều phương pháp hiện đại; các mô hình để học chuyển giao chưa phải là mô hình mới, do đó có thể nâng cấp ứng dụng các mô hình mới hơn; phương pháp sử dụng thuật toán tìm đường biên của viên thuốc, do đó nhiều khả năng sẽ thiếu chính xác trong các trường hợp sự tương phản giữa viên thuốc và nền thấp, dẫn đến hiệu quả thực tế có thể sẽ không chính xác.

(4) Năm 2020, Suiyi Ling và cộng sự [9] cũng có công trình nghiên cứu về nhận dạng viên thuốc, tác giả sử dụng phương pháp mạng nơ-ron tích chập đa luồng; trong đó, để nhận dạng hình dáng loại viên thuốc, sử dụng mô hình mạng W^2-net được xây dựng dựa trên kiến trúc mạng U-net đơn giản; hệ thống được thực nghiệm

trên bộ dữ liệu CURE chứa 8.973 hình ảnh của 196 lớp viên thuốc, Hình 1.12 mô tả tổng quan mô hình của hệ thống.



Hình 1.12. Mô hình hệ thống nhận diện viên thuốc đa luồng CNN

Bảng 1.2. So sánh hiệu quả mô hình Few-shot learning

Bộ dữ liệu	**NIH (%)**	**CURE (%)**
CTM	61,2	50,4
MTL	58,7	47,7
MS	**64,2**	**53,7**

Kết quả, mô hình được nhóm tác giả kiểm tra trên **20%** ảnh viên thuốc trong tập dữ liệu (khoảng 1.716 ảnh) thu được độ chính xác **IoU là 94%** và kết quả toàn bộ hệ thống được trình bày trong **Error! Reference source not found.**. Ưu điểm, hệ thống có khả năng nhận dạng hình dáng, màu sắc và ký tự trên viên thuốc; tác giả thu thập được bộ dữ liệu CURE có số lượng ảnh và lớp viên thuốc tương đối lớn, đa dạng, thích hợp cho huấn luyện các mô hình học sâu có khả năng ứng dụng thực tế; mô hình xây dựng được có độ chính xác khả quan hơn các phương pháp hiện đại, là một trong những

phương pháp hứa hẹn khả năng phát triển. Tuy nhiên, phương pháp tồn tại một số hạn chế như: sử dụng mạng U-net nhỏ (nhỏ hơn 17,5 lần mạng U-net hoàn chỉnh) để phân vùng ảnh viên thuốc giúp tăng hiệu suất xử lý nhưng sẽ làm giảm độ chính xác của hệ thống.

(5) Yu và Chen [25] đã đề xuất một kỹ thuật sử dụng các đặc điểm về màu sắc, hình dạng và dấu ấn của viên thuốc để nhận dạng viên thuốc và đạt được độ chính xác là 97,1% trên bộ dữ liệu hình ảnh thuốc của riêng mình gồm 2.500 viên thuốc. Việc tăng cường dữ liệu được thực hiện bằng cách thay đổi ngẫu nhiên độ sáng, độ tương phản, độ xoay, v.v. của từng hình ảnh viên thuốc thu được; do đó, tạo ra một tập dữ liệu hình ảnh với tổng số 12.500 hình ảnh viên thuốc. Neto và cộng sự [14] đề xuất một máy trích chọn đặc trưng viên thuốc để phân loại thuốc dựa trên hình dạng và màu sắc của viên thuốc đã sử dụng; trình trích xuất đặc trưng được đánh giá bằng cách sử dụng các bộ phân loại KNN, SVM và Bayes; để trích xuất các đặc trưng, họ đã sử dụng hai bộ dữ liệu PILL BR và NIH NLM PIR và đạt được độ chính xác lần lượt là 99,85% và 99,82% trong 0,01006 và 0,0081 giây. Wang và cộng sự [16] đã giới thiệu kỹ thuật học sâu được nhấn mạnh (Highlighted Deep Learning - HDL) để xác định vỉ thuốc, các đặc trưng phân đoạn và mô tả có thể được trích xuất bằng kỹ thuật HDL; kỹ thuật này sử dụng CNN để phân loại đúng loại vỉ và bất biến đối với việc xoay và thay đổi ánh sáng; có độ chính xác gần như 100% trên cơ sở dữ liệu của 272 loại vỉ thuốc.

1.8. Kết chương

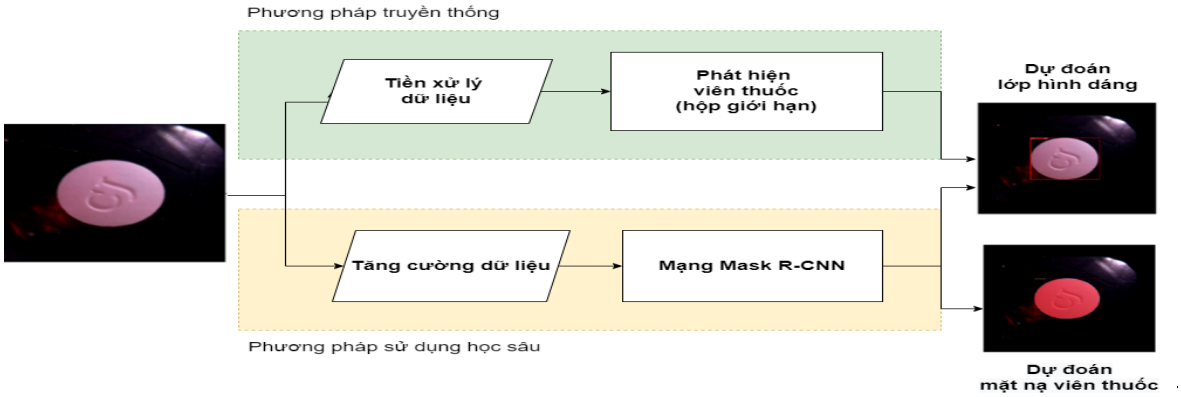
Hiện nay các nghiên cứu liên quan, nhất là các phương pháp dựa trên kỹ thuật tiền xử lý ảnh, tìm cạnh, tìm đường biên... chưa thu được độ chính xác cao trên bộ dữ liệu ảnh thuốc do người dùng cung cấp (vì dữ liệu này có độ tương phản viên thuốc và nền thấp, viên thuốc bị xoay, bị chồng đè, che lấp…), do đó, việc ứng dụng các mô hình học sâu sử dụng kỹ thuật phân đoạn lớp đối tượng (instance segmentation), Unet, Linknet... là một trong những định hướng nghiên cứu tiềm năng hiện nay để giải quyết bài toán.

Qua quá trình khảo sát một số phương pháp, các lý thuyết, vấn đề liên quan bài toán nhận dạng hình dáng loại viên thuốc; luận văn đề xuất phương pháp giải quyết bao gồm các bước: bắt đầu bằng việc tăng cường dữ liệu để làm đa dạng hơn các ảnh viên thuốc mẫu sau đó tiến hành phân tích tập dữ liệu hình ảnh viên thuốc trong bộ dữ liệu CURE; kích thước, hình dạng, màu sắc, dấu ấn... đều là các thông tin tạo nên đặc trưng của viên thuốc được trích xuất thông qua một mạng nơ-ron nhân tạo khung (backbone) học chuyển giao từ mô hình mạng học sâu (như: Resnet50, Resnet101) với trọng số được huấn luyện trước trên bộ dữ liệu COCO kết hợp với kiến trúc mạng Mask R-CNN để huấn luyện mô hình học sâu (model); sau khi tạo cơ sở dữ liệu và mô hình, hệ thống đề xuất có khả năng thực hiện trích xuất đặc trưng tự động của hình ảnh các viên thuốc (theo định dạng và quy định chung trong việc thu thập tập dữ liệu hình ảnh viên thuốc) và sau đó trả về kết quả dự đoán và xác định viên thuốc.

Chương 2: XÂY DỰNG HỆ THỐNG PHÁT HIỆN VÀ NHẬN DẠNG HÌNH DÁNG LOẠI VIÊN THUỐC

2.1. Mô hình hệ thống

Để đánh giá và đề xuất phương pháp hiệu quả giải quyết bài toán, luận văn đã nghiên cứu và thực nghiệm một số kỹ thuật, phương pháp phổ biến hiện nay; trong đó, luận văn đã tập trung vào 02 hướng tiếp cận chính để giải quyết bài toán nhận dạng hình dạng viên thuốc là: ***(1)*** Phân đoạn và phát hiện viên thuốc bằng kỹ thuật truyền thống, trong đó, sẽ tiến hành xử lý ảnh, phát hiện cạnh của các viên thuốc qua đó nhận dạng hình dáng viên thuốc; và ***(2)*** Phát hiện viên thuốc dựa trên kỹ thuật học máy, học sâu.



Hình 2.1. Mô hình tổng quan của hệ thống

2.2. Các tiêu chí đánh giá

Các mô hình phân loại hình ảnh phải được đánh giá để xác định xem chúng hoạt động tốt như thế nào so với các mô hình khác. Dưới đây là một số thông số chính để tính toán tính hiệu quả của các mô hình được sử dụng trong phân loại hình ảnh:

- Precision: Là một số liệu được xác định cho mỗi lớp mà mô hình dự đoán. Độ chuẩn trong một lớp cho chúng ta biết tỷ lệ dữ liệu mà mô hình học máy (ML) dự đoán thuộc về lớp đó đúng là lớp trong dữ liệu xác thực; được tính theo công thức (2.1).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2.1}$$

- TP thể hiện số các kết quả *Đúng tích cực (True Positives)*: Là số lượng mẫu được dự đoán thuộc về một lớp và chính xác thuộc về lớp đó.
- FP đại diện cho *Sai tích cực (False Positive)*: Là số lượng mẫu được dự đoán là thuộc về một lớp trong khi chúng hoàn toàn không thuộc về lớp đó.

- Recall: Recall tương tự như precision được xác định cho mỗi lớp, cho biết tỷ lệ dữ liệu trong tập kiểm tra thuộc về lớp được xác định chính xác (hay dự báo chính xác các mẫu positives thuộc về đúng các lớp positives); được tính theo công thức (2.2).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.2}$$

Trong đó: FN đại diện cho *Sai tiêu cực (False negatives)* là số lượng mẫu mà mô hình dự đoán không thuộc một lớp trong khi chúng thực sự thuộc về lớp đó.

- F1 Score giúp đạt được sự cân bằng giữa độ precision và recall để có được ý tưởng trung bình về cách mô hình hoạt động; được tính theo công thức (2.3).

$$\text{F1 Score} = \frac{2 \text{ x } (Precision \text{ x } Recall)}{Precision + Recall} \tag{2.3}$$

Precision và recall phần lớn phụ thuộc vào vấn đề mà mô hình phân loại đang cố gắng giải quyết. Recall là một số liệu quan trọng, đặc biệt là trong các vấn đề liên quan đến phân tích hình ảnh y tế, như phát hiện viêm phổi từ chụp X-quang ngực, trong đó âm tính giả không thể xuất hiện để ngăn chẩn đoán bệnh nhân khỏe mạnh khi họ thực sự mắc bệnh. Cần phải có độ chính xác để tránh các hiện tượng dương tính giả, như phát hiện spam email; nếu một email quan trọng được phân loại là thư rác, thì người dùng sẽ phải đối mặt với các vấn đề nghiêm trọng.

- IoU (Intersection over Union) là chỉ số đánh giá được sử dụng để đo độ chính xác của mô hình phát hiện đối tượng (trong đó có Mask R-CNN), được tính bằng tỉ lệ của diện tích vùng chồng lên nhau (giữa *predicted bounding box* và *ground-truth bounding box*) và tổng diện mà hai bounding box này đang chiếm. IoU càng gần đến **1** thì mô hình càng chính xác và ngược lại.

- **mAP** (mean Average Precision) là độ đo được sử dụng phổ biến hiện nay cho bài toán Object Detection; là giá trị trung bình của thông số Độ chuẩn trung bình (Average Precision - AP) trên tất cả các lớp với tất cả các giá trị ngưỡng giao hộp gới hạn đối tượng IoU (để xác định các giá trị dự đoán là TP, TN, FP hay FN; IoU càng cao thì dự đoán càng tốt). Thông thường, các mô hình phát hiện đối tượng được đánh giá với các ngưỡng IoU khác nhau, trong đó mỗi ngưỡng có thể đưa ra các dự đoán khác với các ngưỡng khác. Để tính toán mAP, hãy bắt đầu bằng cách tính AP cho mỗi lớp; giá trị trung bình của các AP cho tất cả các lớp là mAP, được tính theo công thức (2.4).

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \quad (2.4)$$

mAP là một thông số tốt để đánh giá mô hình phát hiện đối tượng vì thông qua mối quan hệ giữa Precision và Recall giúp đánh giá độ chính xác phân lớp; và Precision, Recall thay đổi khi ngưỡng IoU thay đổi, do đó tại một giá trị IoU xác định, có thể so sánh độ tốt của mô hình (ví dụ: mAP@0.5 = 80 tức là tại ngưỡng IoU = 0.5 thì độ chính xác của mô hình là 80%).

2.3. Thu thập dữ liệu

Trong lĩnh vực máy học nói chung và bài toán phát hiện, nhận dạng hình dáng loại viên thuốc nói riêng, việc lựa chọn, thu thập bộ dữ liệu huấn luyện và đánh giá đóng vai trò hết sức quan trọng; bộ dữ liệu là thông tin căn bản, nền tảng để huấn luyện, xây dựng bộ trọng số cho mô hình, giúp mô hình "học" được những thông tin hữu ích, có khả năng áp dụng giải quyết bài toán thực tế. Qua nghiên cứu [9] cho thấy, bên cạnh bộ dữ liệu phổ biến NIH NLM của ngân hàng thư viện ảnh thuốc của Mỹ thì bộ dữ liệu CURE do nhóm tác giả Suiyi Ling xây dựng và cung cấp công khai là một trong những bộ dữ liệu đa dạng, chất lượng ảnh sát với thực tế do người dùng cung cấp nhất. Do đó, chúng tôi đã tải và sử dụng bộ dữ liệu CURE do nhóm tác giả cung cấp với kích thước hơn 20GB dữ liệu, chứa **8.973** hình ảnh của **196** loại viên thuốc, mỗi ảnh có kích thước 2044 × 2044 pixel. Bảng 2.1 miêu tả chi tiết thông tin bộ dữ liệu CURE và cho thấy sự hiệu quả của CURE hơn so với bộ dữ liệu NIH NLM.

Bảng 2.1. So sánh bộ dữ liệu NIH và CURE

Đặc điểm	**NIH NLM**	**CURE**
Số lượng ảnh thuốc	7000	8973
Số lượng lớp thuốc	1000	196
Ảnh của mỗi lớp	7	40-50
Điều kiện ánh sáng	1	3
Nền	1	6
Nhãn dấu ấn trên viên thuốc	No	Yes
Nhãn phân đoạn	No	Được gán nhãn 1 phần

2.4. Một số thuật toán phân đoạn và nhận dạng bằng học máy truyền thống

Trước khi nhận dạng, đặc trưng của hình ảnh đã được nâng cao bằng các phương pháp tiền xử lý, chẳng hạn như làm rõ và điều chỉnh cường độ, giúp tăng hiệu quả của việc phân đoạn đối tượng. Phân đoạn được xem xét cho cả các đối tượng chồng chéo và không trùng lặp bằng tất cả các phương pháp. Việc phân đoạn các đối tượng chồng lên nhau theo phương pháp phát triển vùng đã được cải thiện bằng nhiều kỹ thuật, giải thuật khác nhau. Qua nghiên cứu, thực nghiệm, hiện nay có một số phương pháp, kỹ thuật phân đoạn ảnh truyền thống phổ biến sau [25]:

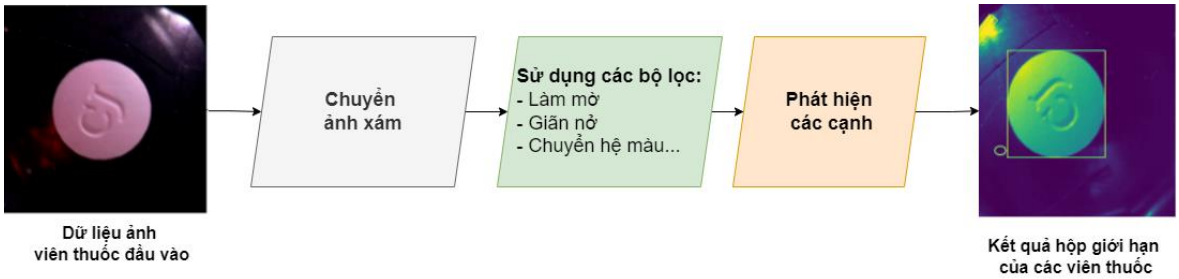
Bảng 2.2. So sánh một số kỹ thuật phân đoạn hình ảnh truyền thống

Kỹ thuật	**Miêu tả**	**Ưu điểm**	**Nhược điểm**
Phương pháp ngưỡng	Tập trung vào việc tìm các giá trị đỉnh dựa trên biểu đồ của hình ảnh để tìm các pixel tương tự	Đơn giản; không yêu cầu tiền xử lý ảnh phức tạp	Nhiều chi tiết có thể bị bỏ qua, thường gặp lỗi ngưỡng
Phương pháp dựa trên cạnh	Dựa trên phát hiện sự gián đoạn không giống như phát hiện tương tự	Tốt cho hình ảnh có độ tương phản cao giữa các đối tượng.	Không thích hợp cho hình ảnh nhiễu
Phương pháp dựa trên khu vực	Dựa trên việc phân vùng hình ảnh thành các vùng đồng nhất	Hoạt động thực sự tốt đối với hình ảnh có lượng nhiễu đáng kể, có thể lấy điểm đánh giá của người dùng để đánh giá nhanh	Tiêu tốn thời gian và bộ nhớ

Các thuật toán phân đoạn truyền thống	Chia hình ảnh thành k cụm đồng nhất, loại trừ lẫn nhau - thu được các đối tượng	Các phương pháp đã được chứng minh, được củng cố bằng logic mờ và hữu ích hơn cho ứng dụng thời gian thực.	Việc xác định hàm chi phí để tối giản có thể khó khăn.
Phương pháp watershed	Dựa trên giải thích cấu trúc liên kết của ranh giới hình ảnh	Các phân đoạn thu được ổn định hơn, các ranh giới được phát hiện có sự khác biệt	Tính toán độ dốc cho các đường biên rất phức tạp.

2.4.1. Kỹ thuật xác định cạnh dựa trên các bộ lọc

Các bộ lọc đóng một vai trò quan trọng như là phương pháp xử lý trước trong phân đoạn hình ảnh. Về nguyên tắc, nhiễu bao gồm các pixel riêng biệt có bề ngoài khác biệt rõ ràng với các pixel liền kề và theo kiến thức này, nhiễu có thể được loại bỏ bằng cách lấy trung bình trong vùng tương tự của dữ liệu hình ảnh thực. Trên thực tế, dữ liệu hình ảnh thực có thể chia sẻ những điểm tương đồng trong các khu vực trung bình này nhưng nhiễu trong các khu vực này thì không. Do đó, quá trình lọc này sẽ giữ dữ liệu hình ảnh thực một cách hiệu quả mà không bị hỏng và giảm nhiễu, Hình 2.2.



Hình 2.2. Xác định cạnh dựa trên các bộ lọc

Qua quá trình sử dụng các bộ lọc chuyển đổi ảnh, lọc nhiễu, phát hiện cạnh có thể làm nổi bật cạnh viên thuốc so với ảnh nền. Kết quả thu được là danh sách các đối tượng có cạnh bao kín, là cơ sở để đánh giá, phân loại hình dạng loại viên thuốc. Tuy nhiên, việc các viên thuốc bị che lấp, đổ bóng, chói sáng… sẽ ảnh hưởng rất lớn tới kết quả phát hiện, do các bộ lọc xử lý ảnh và thuật toán phát hiện cạnh trên chưa thể giải quyết triệt để những thách thức này.

2.4.2. Kỹ thuật xác định bằng biến đổi Wastershed

Để giải quyết thách thức khi các viên thuốc bị che lấp nhau có nhiều cách tiếp cận, trong đó kỹ thuật giãn nở vùng, đặc biệt Watershed được đánh giá là một kỹ thuật phổ biến và hiệu quả. Đây là một dãy thuật toán, được sử dụng trong phân đoạn ảnh, là một công cụ mạnh mẽ dựa trên ranh giới của viên thuốc và tìm các thay đổi cục bộ để phân đoạn hình ảnh [26]. Mô tả đơn giản nhất về biến đổi Watershed xuất phát từ địa lý, theo đó, sườn núi phân chia các khu vực thoát nước bởi các hệ thống sông khác nhau và Watershed là những khu vực thoát nước vào sông hoặc hồ chứa; trong đó các lưu vực là các viên thuốc hoặc khu vực cần xác định. Thuật toán watershed thường được phát triển theo 2 hướng: (1) Theo nguyên lý nước dâng và nó là một hướng tiếp cận truyền thống; (2) Hướng thứ hai được trình bày theo nguyên lý dòng chảy. Luận văn sử dụng phương pháp phân đoạn theo nguyên lý dòng chảy, trong đó, hình ảnh viên thuốc được coi như một bề mặt địa hình với ba loại điểm khác nhau:

 Những điểm chỉ ra cho ta biết đó là điểm tối thiểu

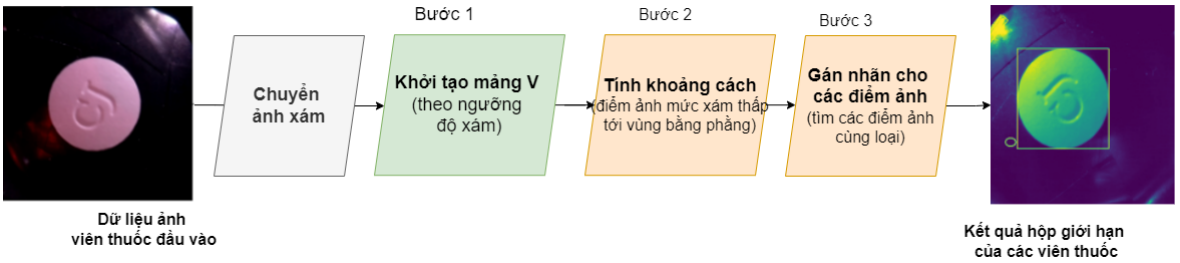
 Những điểm trên sườn dốc, đó là những điểm mà nước chảy vào tối thiểu có xác suất cao nhất.

 Điểm có nước chảy xuống vùng tối thiểu có xác suất cao hơn.

Các nhóm điểm thỏa mãn nhóm thứ hai sẽ được gọi là Catchment basin (tức là vùng chứa các điểm ảnh có chung một tính chất hay còn gọi là lưu vực). Các điểm ảnh thỏa mãn nhóm thứ ba, nơi mà nước rơi xuống và chảy vào nhiều vùng tối thiểu ta gọi đó là các điểm Watershed line (tức là tập hợp các điểm tạo ra đường ngăn cách sự hòa nhập nước, áp dụng vào phân đoạn ảnh có thể gọi nó là đường phân thủy. Ngưỡng chìm được sử dụng để loại bỏ ngọn núi thấp nhất (tương ứng với cạnh yếu nhất trong hình ảnh). Ngọn núi sẽ không được xem xét nếu như chiều cao của chúng nằm dưới ngưỡng này.

Thuật toán watershed dựa trên các thành phần liên thông bao gồm 3 bước: *Bước 1:* Tìm ra các điểm có giá trị độ xám nhỏ hơn giá trị độ xám của các điểm láng

giềng; *Bước 2:* Tìm ra khoảng cách của các điểm ảnh nằm trên vùng bằng phẳng đến điểm ảnh có giá trị xám thấp nhất và gán giá trị cho nó; *Bước 3:* Gán nhãn cho các điểm ảnh có cùng thuộc tính và nhóm nó thành một vùng. Cụ thể mã giả các bước như sau (trong đó, P: biểu diễn một điểm ảnh, f(p): là giá trị xám của điểm ảnh p; N: là điểm ảnh láng giềng của P, f(n) : là giá trị xám của điểm ảnh N; l[p] là mảng được sử dụng để lưu trữ các nhãn. LMAX và VMAX biểu thị giá trị tối đa cho nhãn và khoảng cách tối đa trong hệ thống tương ứng; VMAX xác định khoảng cách giữa pixel đầu tiên của hàng đầu tiên đến pixel cuối cùng của hàng cuối cùng), Hình 2.3.



Hình 2.3. Xác định viên thuốc bằng biến đổi Watershed

Bước 1: Ban đầu, mảng v[p] (là giá trị dùng để lưu khoảng cách từ thấp nhất đến cao nhất của điểm ảnh) được đặt là "0" cho tất cả các phần tử trong mảng. Ảnh viên thuốc đầu vào sẽ được quét từ phía trên cùng từ trái qua phải theo từng hàng. Giá trị v[p] sẽ được đặt là '0' (không đổi) nếu giá trị độ xám của điểm ảnh đang xét thấp hơn hoặc bằng giá trị độ xám của các điểm ảnh lân cận. Các điểm ảnh có giá trị độ xám cao hơn giá trị độ xám của các điểm ảnh lân cận sẽ được đặt là '1'.

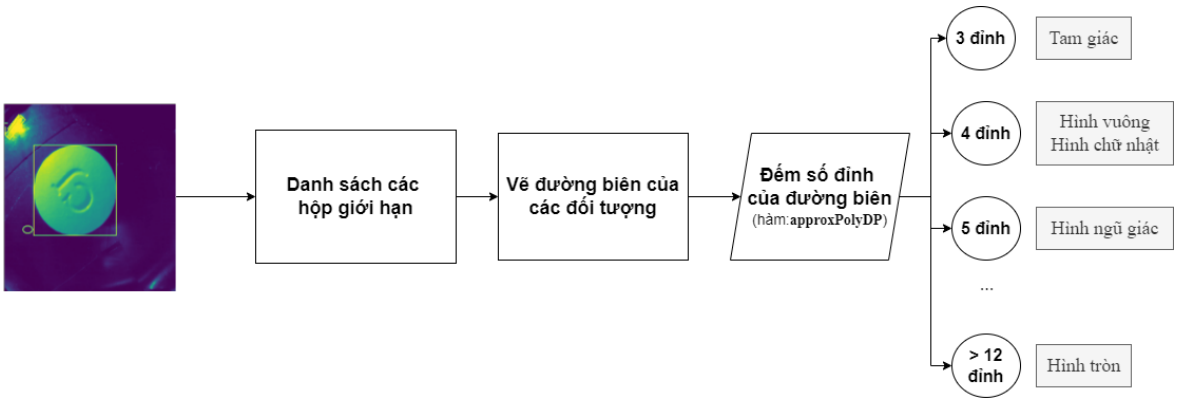
Bước 2 và bước 3: Vùng bằng phẳng là vùng chứa ít nhất 2 điểm ảnh có giá trị độ xám f(p) giống nhau là điểm lân cận nằm cạnh nhau. Trong bước này, chỉ có các điểm ảnh có giá trị v[p] = 0 tìm được ở bước 1 sẽ được xem xét, các điểm ảnh lân cận của điểm ảnh đang được xét phải nằm trên cùng một mặt phẳng giá trị không đổi (Vùng bằng phẳng). Thuật toán duyệt từ trái qua phải theo từng hàng, từ trên xuống dưới. Trong bước này, ảnh sẽ được quét từ trái trên cùng xuống dưới phải. Ban đầu nhãn L(p) của tất cả các điểm ảnh đều được gán = 0. Thuật toán sẽ duyệt ảnh đến khi tất cả các điểm ảnh đều được gán nhãn mới và duyệt đến khi các bước duyệt không có gì thay đổi thì thuật toán dừng lại.

2.5. Nhận dạng hình dáng loại viên thuốc bằng phương pháp truyền thống

Sau khi phân đoạn hình ảnh, cần nhận dạng các viên thuốc từ mỗi hộp giới hạn đối tượng từ bước trên thì cần phân loại xem đấy là hình dạng loại thuốc nào với bao nhiêu phần trăm chắc chắn. Để giải quyết bài toán nhận dạng này, theo phương pháp truyền thống, có hai kỹ thuật: *một là* sử dụng đặc điểm hình học cạnh hoặc đỉnh của đối tượng; *hai là* kỹ thuật đối sánh mẫu với các loại hình viên thuốc mẫu. Cụ thể:

2.5.1. Phương pháp hình học

Dựa trên nhận định hình dáng loại viên thuốc có thể được xác định thông qua việc đếm số đỉnh của hình bao quanh các viên thuốc, ví dụ: số đỉnh bằng 6 là hình lục giác, 5 là ngũ giác, 4 là vuông hoặc chữ nhật, lớn hơn 12 là hình tròn, nhỏ hơn 12 là ellipse..., Hình 2.4.

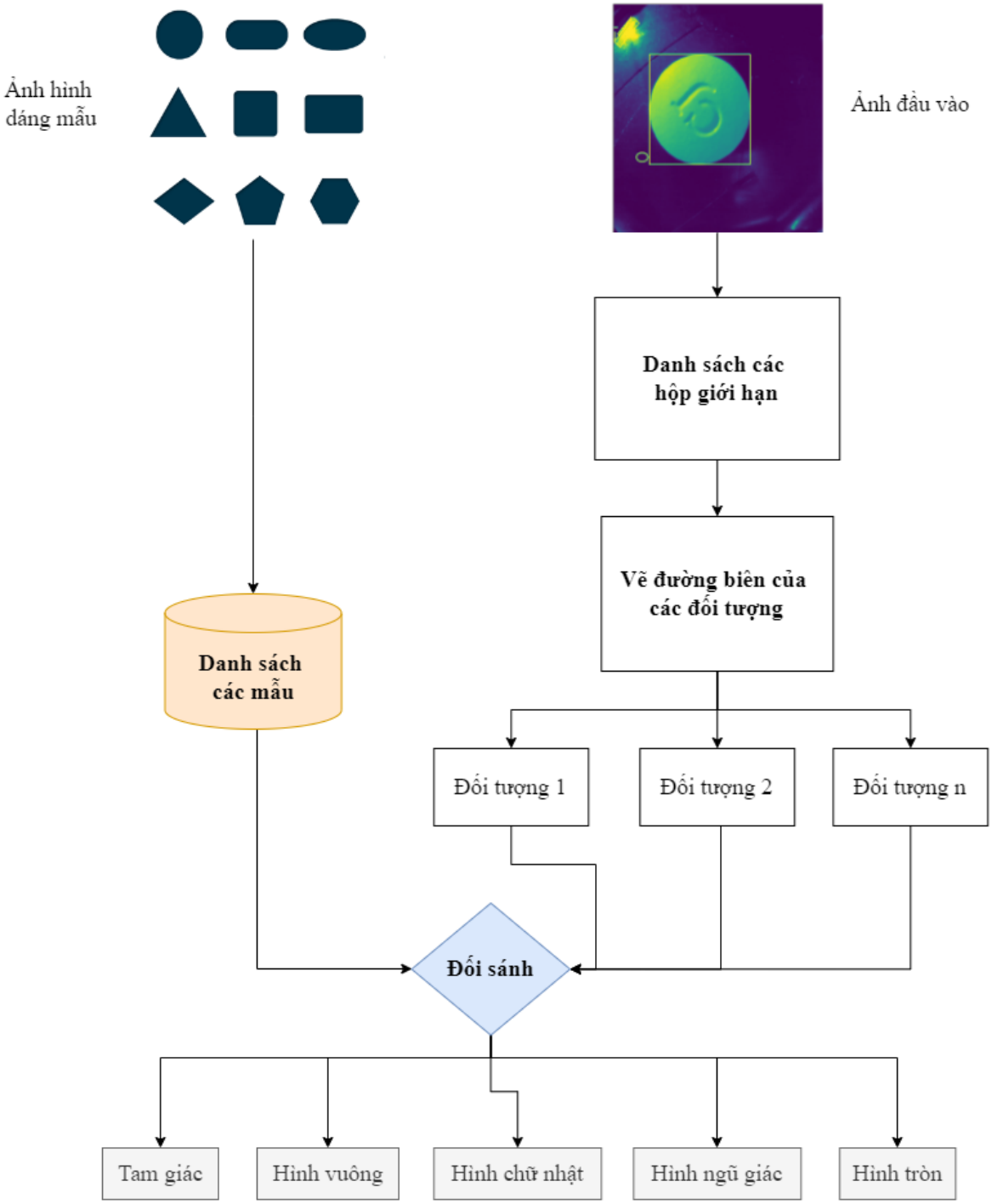


Hình 2.4. Nhận dạng hình dáng bằng phương pháp đếm số đỉnh

2.5.2. Phương pháp đối sánh mẫu

Chúng tôi đã thực nghiệm sử dụng một ảnh mẫu là các loại hình học phổ biến, như: Circle, Elipse, Hexagon… như Hình 2.5 qua đó so sánh ảnh hộp giới hạn thu được với lần lượt từng hình ảnh viên thuốc mẫu theo sơ đồ Hình 2.6.

Hình 2.5. Hình mẫu để sử dụng trong nhận dạng viên thuốc bằng đối sánh mẫu

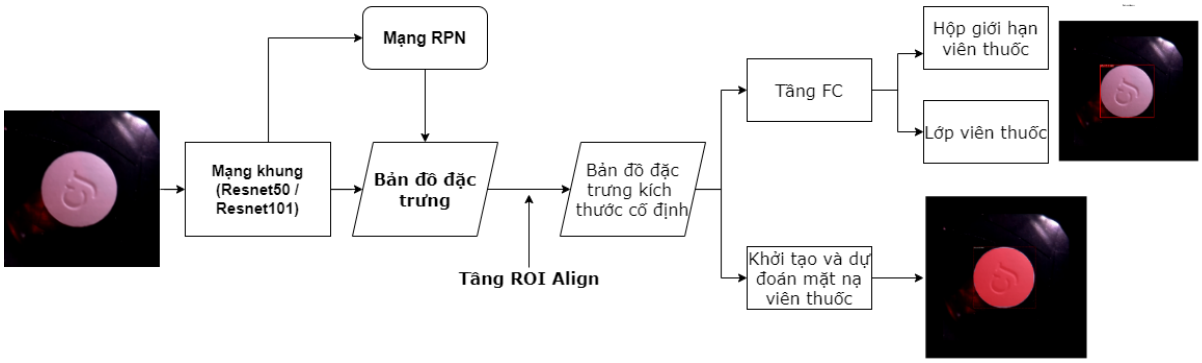


Hình 2.6. Nhận dạng hình dáng viên thuốc bằng kỹ thuật đối sánh mẫu

2.6. Giải pháp phát hiện và nhận dạng hình dáng loại viên thuốc bằng mô hình học sâu Mask R-CNN

2.6.1. Mô hình hệ thống

Nhận dạng hình dáng đối tượng là lĩnh vực phổ biến trong xử lý ảnh và học máy, nhận dạng hình dáng viên thuốc là một bài toán cụ thể của lĩnh vực này. Tuy nhiên, với đặc thù của Tin Sinh học, việc nhận dạng đối tượng đòi hỏi sự chính xác đến từng pixel để hạn chế những dự đoán sai tiêu cực, do đó, phát hiện và nhận dạng hình dáng loại viên thuốc trả về các hộp giới hạn (*bonding boxs)* là chưa đủ độ chi tiết và thiếu tính mở rộng cho hoạt động nghiên cứu bài toán thực tế (ví dụ nhận dạng viên thuốc bị vỡ, sứt góc, quá hạn sử dụng, bay màu...). Để phân đoạn và nhận dạng hình dáng loại viên thuốc, mô hình Mask R-CNN sử dụng bộ dữ liệu được chú thích ở mức pixels, tiến hành một loạt các bước, từ xây dựng Mạng đề xuất vùng RPN và trích xuất các vùng ảnh quan trọng, đến phân lớp vùng đề xuất, tạo, dự đoán mặt nạ của viên thuốc và cuối cùng là làm sạch, phát hiện vị trí, loại viên thuốc. Hình 2.7 mô tả mô hình đề xuất nhận dạng hình dáng loại viên thuốc với Mask R-CNN.



Hình 2.7. Mô hình đề xuất

2.6.2. Tiền xử lý ảnh

Với dữ liệu đầu vào là ảnh thuốc còn nhiều hạn chế, như thiếu độ đa dạng về vị trí, kích thước, độ mờ, điều kiện ánh sáng... của hình ảnh các viên thuốc. Do đó, trước khi huấn luyện mô hình học máy việc tăng cường dữ liệu là rất cần thiết, ảnh hưởng trực tiếp tới độ chính xác của mô hình, nhất là trên các bộ dữ liệu ảnh do người dùng chụp thực tế cung cấp.

Áp dụng trong xây dựng hệ thống giải quyết bài toán, luận văn đã sử dụng thư viện *imgaug* để tăng cường sự đa dạng cho dữ liệu mẫu thông qua các kỹ thuật tiền xử lý ảnh như: lật ảnh (*fliplr)*, áp dụng bộ lọc mờ ngẫu nhiên (*GaussianBlur*), cụ thể:

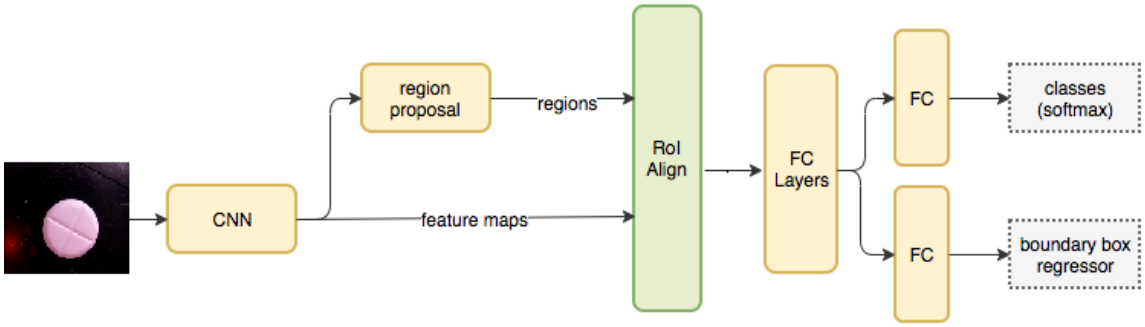
```
import imgaug

augmentation = imgaug.augmenters.Sometimes(0.5, [
                imgaug.augmenters.Fliplr(0.5),
                imgaug.augmenters.GaussianBlur(sigma=(0.0, 5.0))
            ])
```

2.6.3. Phát hiện và nhận dạng bằng Mask R-CNN

Nhìn chung, có thể chia hoạt động của mô hình Mask R-CNN trong phát hiện hình dáng loại viên thuốc thành hai phần: (1) Mạng đề xuất vùng (Region proposal network – RPN) đề xuất các hộp giới hạn khả thi; và (2) Bộ phân lớp mặt nạ nhị phân tạo ra các mặt nạ cho mỗi lớp.

2.6.3.1. Mạng đề xuất vùng khởi tạo các hộp giới hạn viên thuốc

Với dữ liệu đầu vào là các bản đồ đặc trưng của hình ảnh thu được sau khi đi qua mạng nơ-ron tích chập (CNN), cụ thể là mạng ***Resnet50*** hoặc ***Resnet101***, *bước 1:* Mạng đề xuất vùng (RPN) chạy một bộ phân loại nhị phân nhỏ trên nhiều hộp giới hạn (anchor) trên hình ảnh để xác định các Vùng quan tâm (Region of Interest – RoI) và trả về xác suất đánh giá việc có đối tượng hoặc không có đối tượng. Các anchor có điểm đối tượng cao (anchor dương) được chuyển sang giai đoạn hai để được phân loại. Thông thường, ngay cả các anchor dương cũng không bao phủ hoàn toàn các đối tượng. Vì vậy, RPN cũng hồi quy một cách tinh chỉnh (tìm sự cân bằng về vị trí và kích thước) được áp dụng cho các anchor để dịch chuyển nó và thay đổi kích thước từng chút một cho đúng ranh giới của đối tượng; *Bước 2:* ROI xuất ra nhiều hộp giới hạn viên thuốc thay vì một hộp xác định duy nhất và biến đổi chúng thành một chiều cố định; *Bước 3:* Các đặc trưng đã bị biến đổi sau đó được đưa vào các lớp được kết nối đầy đủ (FC) để phân loại bằng cách sử dụng hàm *softmax* và dự đoán hộp ranh giới viên thuốc được tinh chỉnh thêm với thông số IoU bằng cách sử dụng mô hình hồi quy. Hình 2.8 mô tả cụ thể hơn về các bước trên trong dự đoán lớp và hộp giới hạn viên thuốc.



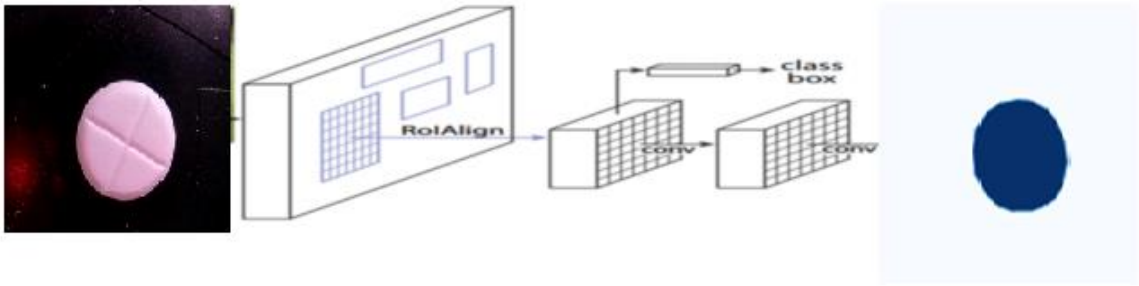
Hình 2.8. Mạng RPN đề xuất các ROI và dự đoán các lớp, hộp giới hạn viên thuốc

Mục tiêu của RPN: Mục tiêu của RPN (RPN Target) là các giá trị huấn luyện cho mạng RPN. Để tạo các mục tiêu, Mask R-CNN bắt đầu với một lưới các anchor bao gồm toàn bộ hình ảnh ở các tỷ lệ khác nhau, sau đó tính giá trị giao nhau của các hộp giới hạn viên thuốc (IoU) của các anchor với viên thuốc thực. Anchor dương là những hộp có IoU> = 0,7 với bất kỳ viên thuốc thực nào và anchor âm là những hộp không bao phủ bất kỳ viên thuốc nào quá 0,3 IoU. Anchor ở giữa (nghĩa là bao phủ một đối tượng bởi IoU > = 0,3 nhưng < 0,7) được coi là trung lập và bị loại khỏi huấn luyện. Để huấn luyện bộ hồi quy RPN, cũng tính toán sự thay đổi kích thước cần thiết để làm cho anchor bao phủ hoàn toàn viên thuốc thực. Đặc biệt, *Tinh chỉnh ROI (ROI Align)* là một đóng góp lớn khác của Mask R-CNN, giải quyết việc tính toán các vùng đề xuất có thể không có cùng kích thước và đưa chúng về cùng kích thước để áp dụng phép nội suy để tính toán các giá trị bản đồ đối tượng tốt hơn, qua đó tăng độ chính xác của Mask R-CNN.

2.6.3.2. Bộ phân lớp mặt nạ viên thuốc và tăng cường dữ liệu

Trong giai đoạn thứ hai, song song với việc để dự đoán độ lệch lớp và hộp giới hạn viên thuốc, Mask R-CNN cũng xuất ra một mặt nạ nhị phân cho mỗi RoI. Nhánh mặt nạ có một đầu ra K_{m^2} chiều cho mỗi RoI, trong đó, mã hóa K mặt nạ nhị phân độ phân giải m × m cho mỗi K lớp. Để áp dụng kỹ thuật này, cần áp dụng hàm *sigmoid* trên mỗi pixel và xác định L_{Mask} là tổn thất entropy chéo nhị phân trung bình (*cross-entropy*). Đối với một RoI được liên kết với lớp ảnh nền (*grouth-truth*) k, L_{Mask} chỉ được xác định trên k-mask tương ứng (đầu ra mặt nạ khác không đóng góp vào giá trị hàm lỗi (*loss*)). Từ công thức trên, hệ thống tính toán được giá trị hàm lỗi của mặt

nạ khi dự đoán viên thuốc, là kết quả rất quan trọng trong phân đoạn viên thuốc ở mức pixel-to-pixel. Hình 2.9 cung cấp mô hình hóa việc Mask R-CNN tạo ra mặt nạ nhị phân cho viên thuốc.

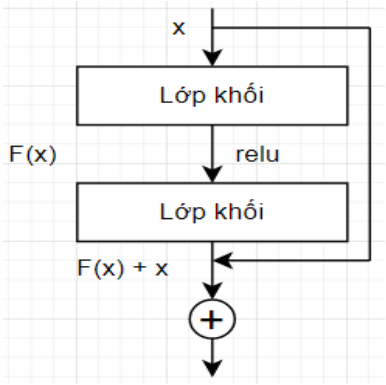


Hình 2.9. Mask R-CNN dự đoán mặt nạ viên thuốc

2.6.4. Huấn luyện mô hình nhận dạng hình dáng viên thuốc

Một vấn đề xảy ra khi xây dựng mạng CNN với nhiều lớp chập sẽ xảy ra hiện tượng *Vanishing Gradient* dẫn tới quá trình học tập không tốt, vậy nên Resnet đã ra đời và giải quyết vấn đề đó.

Cho nên giải pháp mà ResNet đưa ra là sử dụng kết nối "tắt" đồng nhất để xuyên qua một hay nhiều lớp. Một khối như vậy được gọi là một Residual Block (là một chồng các lớp được thiết lập sao cho đầu ra của một lớp được lấy và thêm vào sau một lớp khác nằm sâu hơn trong khối đó), như trong hình sau:



Hình 2.10. Residual Block

ResNet gần như tương tự với các mạng gồm có convolution, pooling, activation và fully-connected layer. Ảnh bên trên hiển thị khối dư được sử dụng trong mạng. Xuất hiện một mũi tên cong xuất phát từ đầu và kết thúc tại cuối khối dư. Hay nói cách khác là sẽ bổ sung Input X vào đầu ra của layer, hay chính là phép cộng mà

ta thấy trong hình minh họa, việc này sẽ chống lại việc đạo hàm bằng 0, do vẫn còn cộng thêm X. Với H(x) là giá trị dự đoán, F(x) là giá trị thật (nhãn), chúng ta muốn H(x) bằng hoặc xấp xỉ F(x). Việc F(x) có được từ x như sau:

$$X \rightarrow \text{weight1} \rightarrow \text{ReLU} \rightarrow \text{weight2}$$

Giá trị H(x) có được bằng cách:

$$F(x) + x \rightarrow \text{ReLU}$$

ResNet sử dụng các kết nối tắt (kết nối trực tiếp đầu vào của lớp (n) với (n+x) được hiển thị dạng mũi tên cong. Qua mô hình nó chứng minh được có thể cải thiện hiệu suất trong quá trình training model khi mô hình có hơn 20 lớp.

Hiện nay, Mask R-CNN đã được các nhà nghiên cứu huấn luyện trên nhiều bộ dữ liệu lớn, trong đó có bộ dữ liệu COCO (là bộ dữ liệu phục vụ các bài toán phát hiện đối tượng, phân đoạn ảnh, chú thích ảnh với hơn 1.5 triệu đối tượng thuộc về 80 lớp khác nhau); do đó, việc sử dụng kỹ thuật học chuyển giao với bộ trọng số được huấn luyện trước trên bộ dữ liệu COCO trên sẽ đem lại hiệu quả cho hệ thống xây dựng, giúp khởi tạo bộ trọng số hợp lý hơn, tăng hiệu quả trích xuất các đặc trưng viên thuốc. Và áp dụng dữ liệu đã được tăng cường, ta có đầu vào là 02 bộ dữ liệu: *dataset_train* và *dataset_val*, sử dụng tỉ lệ học máy *LEARNING_RATE* khởi tạo không quá nhỏ cũng không quá lớn sẽ giúp mô hình huấn luyện hiệu quả hơn (nếu tỉ lệ quá nhỏ thì mô hình lâu hội tụ, nếu tỉ lệ quá lớn thì mô hình huấn luyện thiếu sự chính xác); đồng thời, cần một số lần lặp *epochs* đủ lớn để huấn luyện trên toàn bộ bộ dữ liệu thì mô hình mới thu được độ chính xác cao. Tuy nhiên, việc áp dụng dừng huấn luyện (*callback*) khi độ chính xác đạt đến giá trị nhất định là không thể thiếu, điều này làm giảm khả năng mô hình bị *overfitting* trên bộ dữ liệu huấn luyện. Các thông số tối ưu như Adam, SGD, RmsProp… được áp dụng sẽ giúp cập nhật các bộ trọng số hợp lý hơn cho mô hình.

Kết quả của Mask R-CNN khi dự đoán một ảnh trả về theo dạng từ điển với các khóa cho *hộp giới hạn, mặt nạ, lớp viên thuốc* và *độ chính xác*. Cụ thể:

- ' ***rois*** ': Vùng quan tâm (ROI) cho các viên thuốc được phát hiện.

- ' ***mask*** ': Mặt nạ cho các viên thuốc được phát hiện.
- ' ***class_ids*** ': Số định danh tương ứng với lớp viên thuốc được phát hiện.
- ' ***score*** ': Xác suất tin cậy cho mỗi lớp viên thuốc dự đoán.

2.6.5. Khởi tạo cấu hình mô hình và bộ dữ liệu ảnh thuốc

Để huấn luyện mô hình Mask R-CNN trên bộ dữ liệu ảnh thuốc, việc cấu hình những thông tin chính của mạng và khởi tạo bộ dữ liệu lưu trữ thông tin ảnh thuốc (*PillsConfig*: lớp cấu hình viên thuốc; *PillsDataset*: lớp dữ liệu viên thuốc) là không thể thiếu. Các thông số biến sử dụng chính như:

- Số lớp ảnh thuốc (*number_class = 9*).

- Giá trị ngưỡng tối thiểu để xác định một đối tượng là loại hình dạng viên thuốc hay không (*min-confidence*).

- Tọa độ và nhãn các ảnh viên thuốc đã được gán (*annotation_file.json*).

Ngoài ra, để đánh giá độ chính xác của mô hình, chúng tôi sử dụng giá trị *độ chính xác* (precision) và giá trị tỉ lệ chồng lấp giữa vùng ảnh thực và vùng ảnh dự đoán (*IoU Score*). Đáng chú, giá trị IoU score là giá trị hết sức hiệu quả để đánh giá hoạt động của các mô hình học máy trong phân đoạn ảnh; đồng thời, đây là giá trị mà một số các nghiên cứu liên quan như [9], [28] sử dụng, là cơ sở để so sánh hiệu quả với phương pháp đề xuất.

2.7. Kết chương

Qua nghiên cứu và thực nghiệm xây dựng chương trình, chúng tôi đã tiến hành tiếp cận giải quyết bài toán bằng 02 phương pháp (một là kỹ thuật xử lý ảnh truyền thống và một là kỹ thuật học sâu sử dụng mô hình Mask R-CNN). Có thể thấy, vì hệ thống cần xây dựng làm việc với dữ liệu là ảnh kỹ thuật số nên việc tiền xử lý ảnh là hết sức cần thiết cho mọi phương pháp; đồng thời, các thuật toán có độ phức tạp và hiệu quả khác nhau nên kết quả khi đưa vào áp dụng cho hệ thống sẽ khác nhau và cần đánh giá để lựa chọn phương pháp phù hợp. Mặc dù phương pháp truyền thống đơn giản, áp dụng nhiều thuật toán lọc ảnh, đối sánh ảnh... phổ biến nhưng việc xử lý

các thách thức liên quan bài toán nhiều khả năng sẽ kém hiệu quả, nhất là khi ảnh viên thuốc trùng màu với nền. Bên cạnh đó, phương pháp sử dụng mô hình Mask R-CNN là một trong những mô hình mạng nơ-ron nhân tạo tiên tiến, hiệu quả trong giải quyết bài toán phân đoạn ảnh đối tượng, hứa hẹn sẽ đem lại hiệu quả cao hơn trong giải quyết bài toán đặt ra.

Chương 3: KẾT QUẢ THỰC NGHIỆM VÀ HƯỚNG PHÁT TRIỂN

3.1. Môi trường thực nghiệm và bộ dữ liệu

Chương trình thực nghiệm trên nền tảng Google Colab, với hỗ trợ GPU – Testla 4 (40GB), sử dụng ngôn ngữ Python; chạy huấn luyện và kiểm thử trên bộ dữ liệu CURE chia thành **09 loại hình dạng** khác nhau gồm: *'Capsule', 'Double_round', 'Heart', 'Modified_rectangle', 'Octagon', 'Oval', 'Pentagon', 'Round', 'Triangle'*. luận văn đã gom nhóm và gán nhãn pixel gần **1.700 ảnh** của **09 lớp** hình dạng trên (những viên thuốc chưa gán nhãn chủ yếu là 02 loại hình dạng *Round* và *Capsule* đã được gán nhãn lượng lớn 02 loại thuốc này nên việc gán nhán thêm sẽ không làm tăng hiệu quả của mô hình).

Từ các nghiên cứu liên quan như [12], [15] cho thấy, việc phân chia dữ liệu theo tỉ lệ 80% cho bộ dữ huấn luyện và 20% cho bộ dữ liệu kiểm tra giúp cung cấp đủ dữ liệu cho xây dựng mô hình và đánh giá cập nhật bộ trọng số. Do đó, áp dụng với bài toán đặt ra, luận văn chia tập dữ liệu mẫu thành 02 bộ dữ liệu huấn luyện (**1.284** ảnh) và kiểm tra (**414** ảnh) với tỷ lệ tương ứng là 80% và 20% được thể hiện trong Bảng 3.1 và

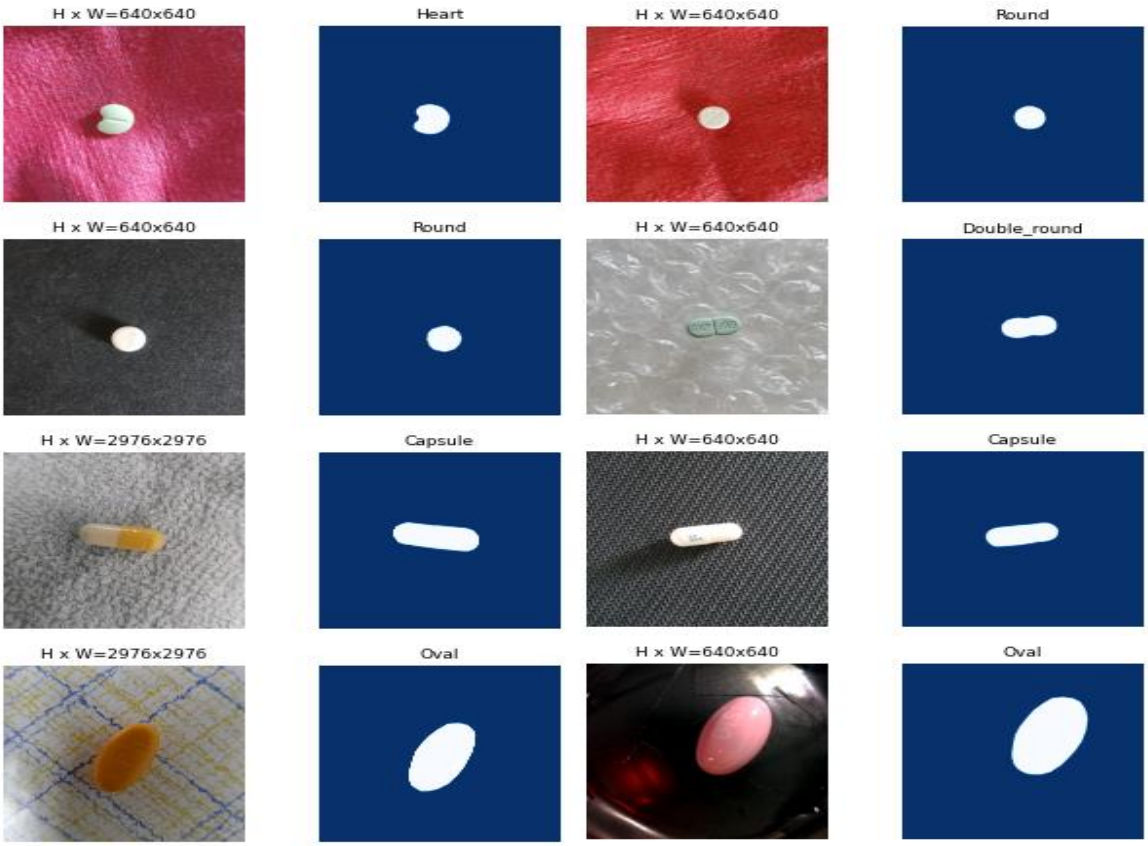
Bảng *3.2*; một số hình ảnh viên thuốc của tập dữ liệu huấn luyện được trình bày trong Hình 3.1.

Bảng 3.1. Số lượng ảnh viên thuốc được gán nhãn của tập huấn luyện

Loại hình dạng	**Số lượng ảnh**	**Loại hình dạng**	**Số lượng ảnh**
1. Capsule	320	2. Double_round	34
3. Heart	30	4. Modified_rectangle	67
5. Octagon	90	6. Oval	228
7. Pentagon	23	8. Round	828
9. Triangle	9		

Bảng 3.2. Số lượng ảnh viên thuốc được gán nhãn của tập kiểm tra

Loại hình dạng	Số lượng ảnh	Loại hình dạng	Số lượng ảnh
1. Capsule	80	2. Double_round	8
3. Heart	8	4. Modified_rectangle	17
5. Octagon	22	6. Oval	57
7. Pentagon	6	8. Round	207
9. Triangle	3		

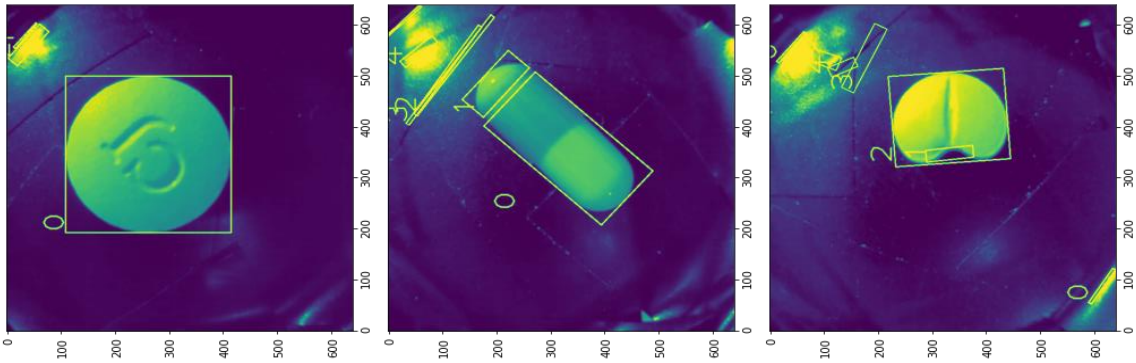


Hình 3.1. Một số mẫu dữ liệu viên thuốc được chú thích pixel

3.2. Kết quả thực nghiệm

3.2.1. Nhận dạng hình dáng viên thuốc bằng phương pháp phát hiện cạnh

Luận văn đã thực nghiệm một số kỹ thuật áp dụng các bộ lọc khác nhau, qua đó các cạnh của viên thuốc có thể được thể hiện rõ hơn, có sự tương phản nhất định so với nền. Tuy nhiên, do ảnh hưởng của các điều kiện tự nhiên (ánh sáng, độ chói, bóng, tương phản…) trên ảnh của bộ dữ liệu mẫu CURE khiến phương pháp truyền thống không thể phát hiện hoàn chỉnh đủ cạnh, viền tại một số viên thuốc hoặc phát hiện sai các nhiễu. Hình 3.2 cho thấy kết quả phát hiện viên thuốc với độ chính xác không cao chứa nhiều kết quả bị nhiễu của phương pháp phát hiện cạnh bằng các bộ lọc và tiền xử lý ảnh.



Hình 3.2. Kết quả phân đoạn trên dữ liệu mẫu bằng phát hiện cạnh bằng tiền xử lý ảnh (openCV)

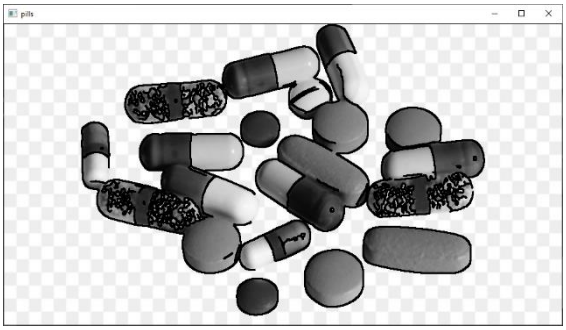
Chúng tôi thực nghiệm nhận dạng hình dáng loại viên thuốc trên bộ dữ liệu kiểm tra chứa **414 ảnh**, độ chính xác đạt **34.71%**; trong đó, do việc dựa vào số đỉnh để nhận dạng hình dáng thiếu sự chi tiết (không phân biệt được một số lớp giống nhau, như: *Oval* và *Modified_rectangle* với *Capsule; Heart* và *Double_round* với *Round*), vì vậy chúng tôi đã gộp các lớp này lại với nhau. Hình 3.3 là một số kết quả nhận dạng hình dáng viên thuốc bằng phương pháp xử lý ảnh truyền thống (sử dụng các bộ lọc và phát hiện cạnh bằng OpenCV).

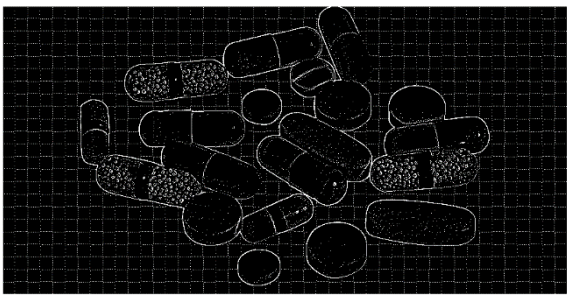
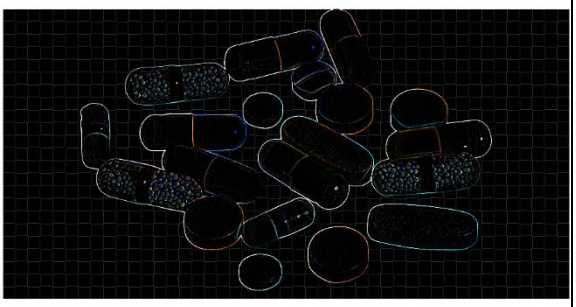
Thời gian xử lý 1 ảnh xấp xỉ 0.43 giây; không mất thời gian xây dựng, huấn luyện mô hình.

Hình 3.3. Kết quả nhận dạng hình dáng loại viên thuốc bằng OpenCV

Ngoài ra, luận văn đã thực nghiệm trên một số ảnh thực tế người dùng cung cấp, khi ảnh có nhiều viên thuốc, nhất là có những viên thuốc có sự chồng, lấp thì phương pháp truyền thống hoạt động không hiệu quả ngay từ bước phân đoạn viên thuốc; không thể phát hiện chính xác cạnh của các viên thuốc dẫn đến không thể phân loại hình dạng theo yêu cầu bài toán đặt ra, kết quả được mô tả cụ thể tại Bảng 3.3.

Bảng 3.3. Kết quả phân đoạn hình ảnh viên thuốc bằng các bộ lọc

Ảnh đã được xám hóa	**Ảnh phân đoạn**
1. Sử dụng OpenCV	
	

2. Sử dụng thư viện PILLOW	
	
3. Sử dụng phương pháp Watershed	
	

3.2.2. Thực nghiệm phát hiện và nhận dạng hình dáng viên thuốc bằng mô hình Mask R-CNN

3.2.2.1. Gán nhãn dữ liệu

Gán nhãn dữ liệu viên thuốc đề cập đến công việc phân loại thủ công hoặc bán tự động và chuyển đổi dữ liệu bằng cách sử dụng các công cụ gán nhãn để xây dựng bộ dữ liệu khớp với yêu cầu của mô hình học sâu. Để phát hiện viên thuốc, hình ảnh huấn luyện và tọa độ vị trí của viên thuốc tương ứng với mỗi hình ảnh bắt buộc phải được xác định một cách cụ thể, chính xác. Mask R-CNN yêu cầu tọa độ đa giác thể hiện hình dạng của viên thuốc và tọa độ vị trí của chúng. Để tạo tọa độ đa giác này, cần sử dụng công cụ gán nhãn hình ảnh để hiển thị tọa độ đa giác và tên lớp cho từng viên thuốc trong ảnh. Chúng tôi đã sử dụng công cụ nền tảng web online là Makesense.ai để gán nhãn pixel cho bộ dữ liệu huấn luyện và đánh giá.

Đây là công việc sử đòi hỏi thời gian và công sức đáng kể. Kết quả gán nhãn thu được các tệp JSON lưu trữ thông tin về lớp hình dạng, tọa độ, ví trí của viên thuốc

cũng như mặt nạ của chúng (*'Capsule', 'Double_round', 'Heart', 'Modified_rectangle', 'Octagon', 'Oval', 'Pentagon', 'Round', 'Triangle'*).

3.2.2.2. Các giá trị tham số huấn luyện mô hình

Qua nghiên cứu, tham khảo một số công trình liên quan như [11], [15], để huấn luyện mô hình theo phương pháp đề xuất, chúng tôi đã thực nghiệm huấn luyện mô hình với các giá trị tham số khởi tạo được trình bày cụ thể tại Bảng 3.4, trong đó:

- Sử dụng thuật toán tối ưu tham số Stochastic Gradient Descent (SGD) là thuật toán tối ưu hóa cơ bản theo họ gradient;
- Tỉ lệ học khởi tạo ban đầu là 0.001;
- Kích thước dữ liệu mỗi lần mô hình học (batch size) khởi tạo bằng 1 là do kích thước ảnh viên thuốc đầu vào lớn và không thay đổi kích thước nhằm giữ đầy đủ thông tin của ảnh (trong đó, ảnh lớn nhất kích thước là 2448 x 2448 pixel) trong khi cấu hình phần cứng của môi trường thực nghiệm là Google Colab pro bị hạn chế bởi cấu hình GPU (40GB).
- Qua thực nghiệm với các số lần lặp khác nhau, để đảm bảo độ chính xác của mô hình và thời gian, hiệu suất xử lý, chúng tôi đã chọn tham số số lần lặp epochs bằng 80.
- Hàm lỗi của mô hình Mask R-CNN là sự kết hợp của hai nhiệm vụ : *một là* mạng FCN để dự đoán hộp giới hạn viên thuốc ; *hai là* nhiệm vụ dự đoán mặt nạ của viên thuốc.

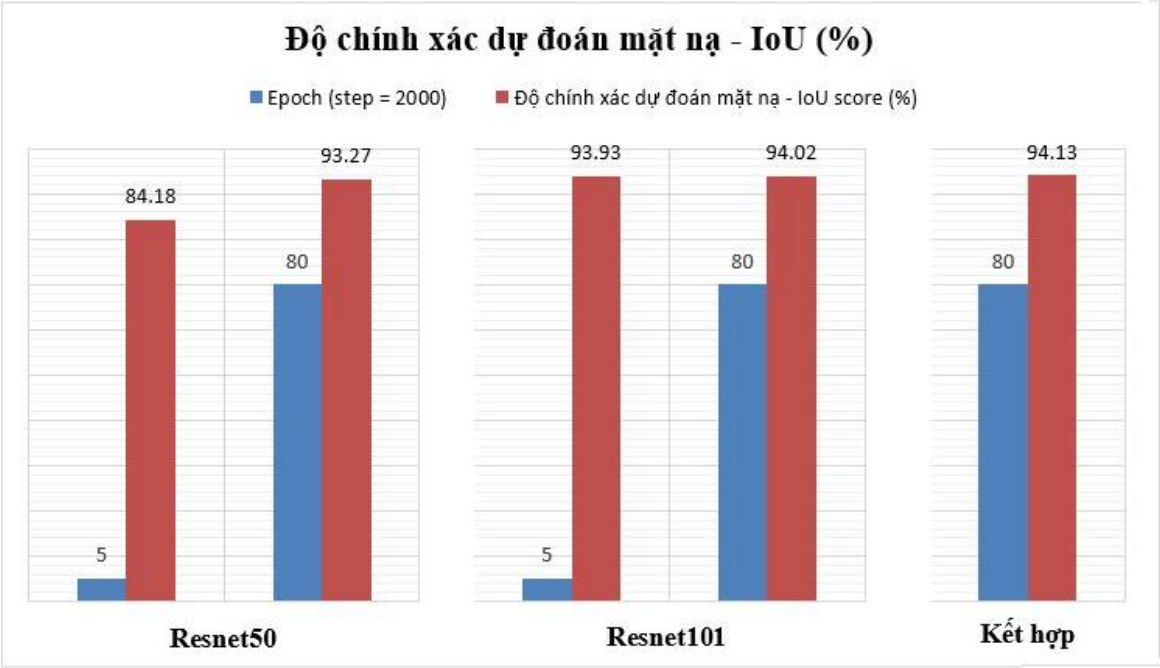
Bảng 3.4. Giá trị tham số huấn luyện mô hình Mask R-CNN

STT	**Tham số**	**Giá trị**
1	*Optimizer*	SGD
2	*learning_rate*	0.001
3	*batch_size*	1
4	*epoch*	80

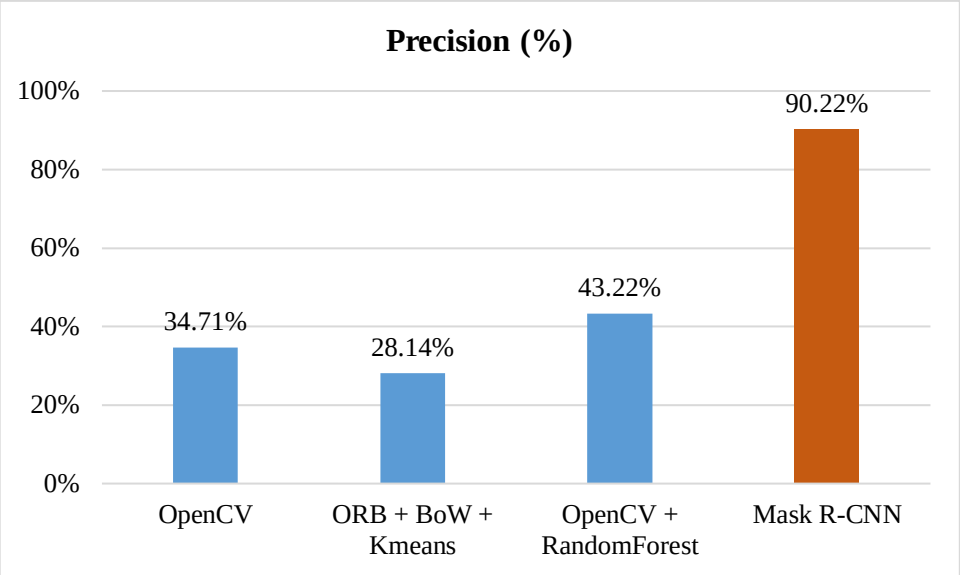
STT	Tham số	Giá trị
5	*loss_function*	rpn_class_loss = 1 rpn_bbox_loss = 1 mrcnn_class_loss = 1 mrcnn_bbox_loss = 1 mrcnn_mask_loss = 1

3.2.2.3. Độ chính xác dự đoán của mô hình

Qua kết quả thực nghiệm trên bộ dữ liệu đánh giá, chúng tôi thu được kết quả độ chính xác khi dự đoán mặt nạ viên thuốc IoU score đạt **94,13%**, độ chính xác phân lớp đạt **90.22%** cao hơn khi áp dụng các phương pháp khác như: OpenCV, ORB + BoW + KMeans, OpenCV + RandomForest; được thể hiện dưới dạng đồ thị ở các đồ thị Hình 3.4 và Hình 3.5. Trong đó, để đánh giá kết quả, chương trình thực nghiệm trên 02 kiến trúc là Resnet50, Resnet101 với số lần lặp khác nhau là 5 lần và 80 lần.



Hình 3.4. Kết quả độ chính xác theo tỉ lệ chồng lấp IoU của mô hình thực nghiệm với các kiến trúc và thông số khác nhau



Hình 3.5. Kết quả so sánh độ chính xác của phương pháp đề xuất với phương pháp truyền thống

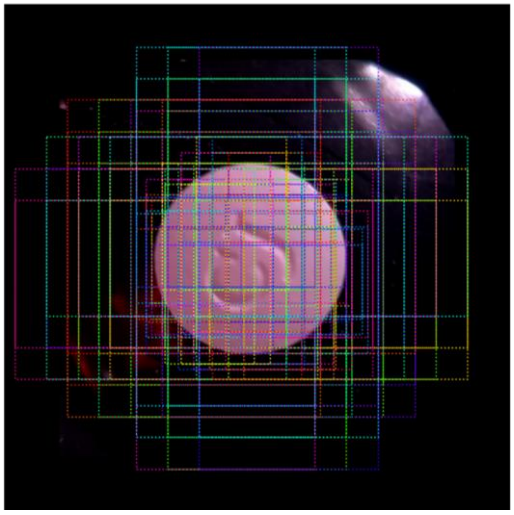
Đặc biệt, luận văn đã so sánh kết quả độ chính xác dự đoán hình dạng loại viên thuốc theo tỉ lệ IoU của các mặt nạ viên thuốc với các nghiên cứu liên quan. Kết quả cho thấy, phương pháp đề xuất có độ chính xác cao nhất, Bảng 3.5 thể hiện kết quả so sánh.

Bảng 3.5. So sánh độ chính xác của mô hình với một số nghiên cứu liên quan

STT	**Phương pháp**	**Độ chính xác IoU (%)**
1.	Phương pháp theo tài liệu số 27 [27]	90
2.	Phương pháp theo tài liệu số 28 [28]	78
3.	Phương pháp theo tài liệu số 15 [15]	94
4.	**Phương pháp đề xuất**	**94.13**

**** Kết quả dự đoán trên 01 mẫu ngẫu nhiên cụ thể như sau:***

(1) Đề xuất vùng chứa viên thuốc

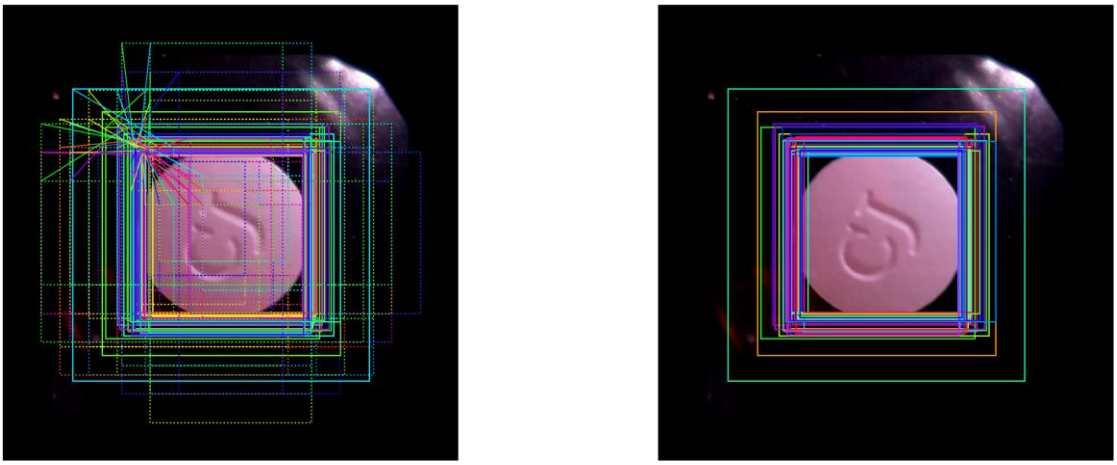


Hình 3.6. Các anchor dương trước khi sàng lọc (chấm) và sau khi sàng lọc (liền)

Hình 3.7. Các anchor box được tinh chỉnh sau khi loại bỏ những hộp độ chính xác thấp

(2) Phân loại vùng đề xuất

Sau khi thu được các vùng đề xuất, hệ thống tiến hành phân loại những vùng thu được này. Chạy trình phân loại đứng đầu trên các đề xuất để tạo ra các đề xuất của lớp và tiến hành thuật toán hồi quy trên các hộp giới hạn.



Hình 3.8. Hiển thị các vùng đề xuất cuối cùng

Hình 3.9. Phân loại các vùng đề xuất hình dạng viên thuốc

(3) Tạo ra các mặt nạ viên thuốc và dự đoán lớp hình dáng

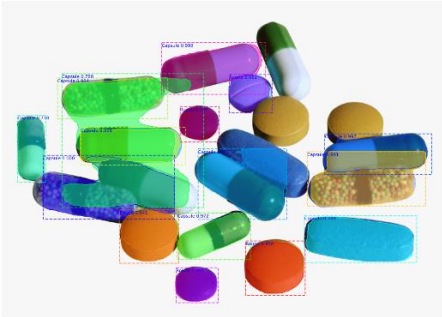
Bước này lấy các vùng phát hiện được (tinh chỉnh hộp giới hạn và định danh lớp) từ bước trước và chạy tầng đầu của Mask để tạo mặt nạ phân đoạn cho các trường hợp.

Hình 3.10. Tạo ra mặt nạ phân đoạn cho các viên thuốc

Hình 3.11. Kết quả phát hiện và nhận dạng viên thuốc

* Ngoài ra, mặc dù bộ dữ liệu CURE là dữ liệu hình ảnh phức tạp, phần lớn là ảnh do người dùng chụp với sự thay đổi của nền, ánh sáng ... và mô hình chúng tôi xây dựng đã huấn luyện và tính toán độ chính xác trên dữ liệu này. Tuy nhiên những ảnh này mới chỉ là ảnh có một viên thuốc, mà trong điều kiện thực tế thường xuất hiện tình huống dữ liệu hình ảnh chứa nhiều viên thuốc và chồng, lấp nhau. Do đó, chúng tôi kiểm tra hiệu quả của mô hình trên một số hình phức tạp hơn, khi mà có nhiều viên thuốc cùng xuất hiện trong ảnh, để cho thấy khả năng ứng dụng vào thực tế của mô hình đề xuất.

Bên cạnh đó, do không có bộ dữ liệu hình ảnh viên thuốc phức tạp làm chuẩn hay công trình nghiên cứu tương tự kiểm thử trên tiêu chí này nên chúng tôi chưa thể kiểm tra trên số lượng lớn để đánh giá độ chính xác. Hình 3.12 là kết quả áp dụng mô hình trên một ảnh phức tạp; trong đó, đã phát hiện và nhận dạng tương đối chính xác nhiều ảnh viên thuốc có sự chồng, lấp nhau, chỉ có một số ít các viên thuốc phức tạp hệ thống chưa thể nhận dạng được.



Hình 3.12. Kết quả phát hiện và nhận dạng hình dáng viên thuốc trên ảnh thực tế bằng Mask R-CNN

3.2.3. Thời gian xử lý

Thời gian huấn luyện mô hình Mask R-CNN sấp xỉ 8 giờ.

Thời gian trung bình để phát hiện và nhận dạng hình dáng loại viên thuốc trong 1 ảnh là khoảng 1.75 giây.

3.3. Kết chương

Từ kết quả thực nghiệm trên cho thấy, các phương pháp phân đoạn dựa trên việc phát hiện cạnh (phương pháp truyền thống) chưa đưa được kết quả có độ chính xác cao, khó có thể áp dụng được cho thực tiễn do ảnh các viên thuốc có thể có màu trắng và màu trùng với màu nền, làm giảm độ tương phản của đối tượng với nền. Đồng thời, các viên thuốc có khả năng bị đổ bóng, chói sáng, chồng, lấp hoặc che nhau, do đó cần có phương pháp tương tự watershed nhưng hiệu quả cao hơn để phân đoạn. Do đó, việc ứng dụng mô hình mạng nơ-ron nhân tạo Mask R-CNN để giải quyết bài toán là hết sức cần thiết, kết quả thực nghiệm đã cho thấy tính khả thi của phương pháp đề xuất trong giải quyết bài toán phát hiện và nhận dạng hình dáng loại viên thuốc. Cụ thể:

3.3.1. Ưu điểm

Phương pháp nhận dạng hình dáng loại viên thuốc dựa trên mô hình học sâu Mask R-CNN với bộ dữ liệu viên thuốc CURE đã thu được kết quả tương đối cao, tỉ lệ dự đoán mặt nạ viên thuốc IoU đạt 94,13%, cho thấy hướng phát triển khả quan khi áp dụng vào bài toán nhận dạng viên thuốc trên các ảnh thực tế do người dùng cung cấp. Một số ưu điểm nổi bật mà luận văn đã đóng góp như sau:

- Trên bộ dữ liệu này, phương pháp đề xuất đã thu được kết quả thực nghiệm với độ chính xác đạt **94.13%** cao hơn phương pháp của *Suiyi Ling và cộng sự năm 2020* [9] thực nghiệm mô hình đa luồng và một số phương pháp khác.

- Đã chú thích mức độ pixel cho bộ dữ liệu hình ảnh viên thuốc với hơn **1.700** viên thuốc, là tiền đề rất lớn cho các bài toán phân đoạn và nhận dạng viên thuốc chính xác cao trong tương lai.

- Qua thực nghiệm cho thấy, phương pháp dựa trên hình học trong nhận dạng hình dáng viên thuốc chưa thực sự hiệu quả, chịu ảnh hưởng lớn từ kết quả tìm đường

bao quanh đối tượng, khó liệt kê hết và chính xác các hình, đặc biệt, không chính xác khi các viên thuốc có hình đặc biệt như trái tim, hoa… hay cạnh viền không được làm nét, mịn; đồng thời, phương pháp nhận dạng hình dáng loại viên thuốc bằng đối sánh mẫu cũng chịu ảnh hưởng lớn từ việc phân đoạn ảnh bằng các bộ lọc tiền xử lý truyền thống nên việc áp dụng cho bài toán thực tế trong đời sống sẽ kém hiệu quả.

3.3.2. Một số hạn chế

Bên cạnh những kết quả đã đạt được, phương pháp đề xuất còn tồn tại một số khó khăn, hạn chế, như:

- Mặc dù chương trình đã được huấn luyện trên những lớp hình dạng viên thuốc phổ biến nhất hiện nay, tuy nhiên, độ đa dạng về hình dáng loại viên thuốc chưa cao (hiện có 09 lớp hình dáng, thiếu nhiều lớp viên thuốc hình dáng đặc biệt), khi áp dụng vào thực tiễn cần bổ sung thêm nhiều dữ liệu mẫu và chú thích điểm ảnh.

- Một số lớp hình dạng loại viên thuốc có số lượng hình ảnh hạn chế, do đó hiệu quả của mô hình khi nhận dạng những viên thuốc có hình dạng này chưa cao.

- Nền tảng áp dụng mô hình Mask R-CNN cần được cải tiến, ví dụ như sử dụng các thư viện Pytorch, Tensorflow... mới hơn để tăng hiệu năng, hiệu quả của chương trình.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TIẾP

Mặc dù ngay nay nhiều phương pháp nhận dạng đối tượng nói chung đã ra đời, đạt được hiệu quả, hiệu suất cao, nhưng nhận dạng viên thuốc vẫn luôn là trọng tâm nghiên cứu do tính ứng dụng và thiết yếu trong đời sống của nó. Nhận dạng hình dáng viên thuốc là một bài toán đang phát triển thành một hệ thống ứng dụng công nghệ thông tin trong chăm sóc sức khỏe thường ngày của con người. Do đó, việc ứng dụng công nghệ nhận dạng hình dáng loại viên thuốc phục vụ các cơ sở y tế, người dùng thuốc viên là một phương pháp khả thi, góp phần phục vụ hiệu quả nhu cầu thực tế của đời sống.

Trong phạm vi nghiên cứu, luận văn đã trình bày bài toán nhận dạng hình dáng loại viên thuốc; trong đó, tập trung vào một số kỹ thuật nhằm giải quyết bài toán, như: phân đoạn viên thuốc bằng phương pháp tìm đường biên truyền thống; phân đoạn và nhận dạng hình dáng viên thuốc bằng mô hình học sâu. Từ đó đề xuất xây dựng chương trình thực nghiệm thông qua phương pháp sử dụng mô hình Mask R-CNN, xây dựng mô hình huấn luyện của hơn **1.700** hình ảnh viên thuốc thuộc **09** loại hình dáng khác nhau, từ đó thu được độ chính xác nhận dạng đạt xấp xỉ **94,13%**. Việc nghiên cứu, phát triển giải pháp nhận dạng hình dáng viên thuốc bằng mô hình học sâu một cách tự động để thay thế cho nhận dạng bằng phương pháp truyền thống hoặc yêu cầu người dùng tự mô tả thụ động là một hướng nghiên cứu mới, đáp ứng được tình hình thực tế, nhu cầu hiện nay của nhiều nước, trong đó có Việt Nam, đặc biệt là trong kỷ nguyên khoa học công nghệ bùng nổ, việc trang bị các thiết bị quay phim, ghi hình, chụp ảnh cá nhân ngày càng phổ biến và hiện đại.

Bám sát mục tiêu, nhiệm vụ, sử dụng đúng đắn các phương pháp nghiên cứu khoa học, luận văn đã thu được một số thành công và về cơ bản đã đạt được mục tiêu, nhiệm vụ đặt ra. Trong tương lai, để cải thiện hiệu suất của hệ thống, luận văn định hướng mở rộng nghiên cứu tăng cường bộ dữ liệu mẫu, giảm nhiễu; đồng thời, cần xem xét, nghiên cứu ứng dụng những thành công của các phương pháp tiến tiến hiện đại bắt nguồn từ thị giác máy tính, khai thác dữ liệu lớn và khoa học máy tính, ứng dụng các kỹ thuật mới để phát hiện, nhận dạng đối tượng.

Hướng phát triển tiếp: Việc phát hiện và nhận dạng hình dáng loại viên thuốc là nền tảng hết sức quan trọng cho lĩnh vực Tin Sinh học, đặc biệt là vấn đề nghiên cứu phân loại viên thuốc. Do đó, để cải tiến hiệu quả, trong tương lai có thể nghiên cứu các kỹ thuật học sâu mới hơn, nâng cấp, cải thiện thuật toán Mask R-CNN; đồng thời, nghiên cứu, phát triển việc phân loại viên thuốc dựa trên màu sắc, ký hiệu viên thuốc, quá đó giúp đưa việc nhận dạng viên thuốc trở thành hệ thống hoàn chỉnh hơn, không chỉ dừng lại ở bài toán nhận dạng hình dáng loại viên thuốc.

Luận văn là công trình nghiên cứu công phu, nghiêm túc, song do đây là vấn đề khó và phức tạp, phạm vi nghiên cứu rộng, cộng thêm những khó khăn khách quan, cũng như kiến thức còn hạn chế nên chắc chắn còn nhiều khiếm khuyết. Rất mong nhận được sự quan tâm, góp ý của các nhà khoa học, nhà hoạt động thực tiễn và đồng nghiệp. Cuối cùng, xin chân thành cảm ơn các đơn vị liên quan, các đồng chí, đồng nghiệp, đặc biệt là thầy hướng dẫn khoa học đã tận tình giúp đỡ để hoàn thành luận văn này./.

TÀI LIỆU THAM KHẢO

[1] A. Jacobson, "Medication errors statistics 2022," *SingleCare Team,* 01/2022.

[2] Tariq, Rayhan A.; Vashisht, Rishik; Sinha, Ankur; Scherbak, Yevgeniya, "Medication Dispensing Errors And Prevention," *NCBI,* 01/2021.

[3] Kapil G Zirpe, Bhavika Seta, Sharvari Gholap, "Incidence of Medication Error in Critical Care Unit of a Tertiary Care Hospital: Where Do We Stand?," *PMC,* 2020.

[4] Mandal, Manav, "Introduction to Convolutional Neural Networks (CNN)," Data Science Blogathon., 2021.

[5] Girshick, Ross, "Fast r-cnn.," volume 10.1109/ICCV.2015.169. , 2015.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *arXiv,* 2016.

[7] R. Girshick, "Mask R-CNN," *Facebook AI Research (FAIR),* 2018.

[8] Hyuk-Ju Kwon, Hwi-Gang Kim, Sung-Hak Lee, "Pill Detection Model for Medicine Inspection Based on Deep Learning," *chemosensors - MDPI,* 2021.

[9] Suiyi Ling et al, "Few-Shot Pill Recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , volume doi: 10.1109/CVPR42600.2020.00981, pp. 9786-9795, 2020.

[10] World Health Organization, Medication Errors, ISBN 978-92-4-151164-3, 2016.

[11] A. Woodbury, "Increasing Medication Safety with Deep Learning Image Recognition," *RxVision,* 2020.

[12] Wong, Y. F. et al., "Development of fine-grained pill identification algorithm using deep convolutional network," *J. Biomed. Inform. 74,* 2017.

[13] S. Tangwattananuwat, "The Identification of Pill Images Using Convolutional," 2020.

[14] M.A.V. Neto, J.W.M. de Souza, P.P. Reboucas Filho and W.D.O. Antonio, "CoforDes: An invariant feature extractor for the drug pill identification," *IEEE 31st Int. Symp. on Computer-Based Medical Systems (CBMS), Karlstad,* 2018.

[15] N. L. o. Medicine, "nlm.nih.gov," 2016. [Trực tuyến]. Available: https://www.nlm.nih.gov/pubs/techbull/ma16/brief/ma16_pill_challenge.html.

[16] J.S. Wang, A. Ambikapathi, Y. Han, S.L. Chung, H.W. Ting and C.F. Chen, "Highlighted deep learning based identification of pharmaceutical blister packages," *IEEE 23rd Int. Conf. on Emerging Technologies and Factory Automation (ETFA), Turin,* 2018.

[17] Maier, Andreas & Syben, Christopher & Lasser, Tobias & Riess, Christian, "A Gentle Introduction to Image Segmentation for Machine Learning," *A gentle introduction to deep learning in medical image processing. Zeitschrift für Medizinische Physik.*, volume 29, 10.1016/j.zemedi.2018.12.003. , 2021.

[18] N. Barla, "The Complete Guide to Panoptic Segmentation," PerceptronAI, https://www.v7labs.com/blog/panoptic-segmentation-guide, 2022.

[19] Arpan Kumar, Anamika Tiwari , "A Comparative Study of Otsu Thresholding and K-means Algorithm of Image Segmentation," *International Journal of Engineering and Technical Research (IJETR),* volume 9, number 5, p. 1, 2019.

[20] H. Bandyopadhyay, "An Introduction to Image Segmentation: Deep Learning vs. Traditional," V7, 2022.

[21] Michielan, Lisa & Terfloth, Lothar & Gasteiger, Johann & Moro, Stefano, "Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates.," *Journal of chemical information and modeling,* volume 49. 10.1021/ci900299a., pp. 2588-605, 2009.

[22] Gershenson, Carlos, "Artificial Neural Networks for Beginners," 2003.

[23] Radhamadhab Dalai, Kishore Kumar Senapati, "Comparison of Various RCNN techniques for Classification of Object from Image," *International Research Journal of Engineering and Technology (IRJET),* volume 04, 2017.

[24] N. Usuyama, L. Naoto, "ePillID Dataset: A Low-Shot Fine-Grained Benchmark for Pill Identification," *arXiv,* 2020.

[25] Chen, J. Yu and Z., "Accurate system for automatic pill recognition using imprint information," *IET Image Process.*, volume 9, 2015.

[26] S. Prasad, "analytixlabs," 2022. [Trực tuyến]. Available: https://www.analytixlabs.co.in/blog/what-is-image-segmentation/. [acess 2022].

[27] K V, Lalitha & .R, Amrutha & Michahial, Stafford , "Implementation of Watershed Segmentation," *IJARCCE,* volume 5, number 10.17148/IJARCCE.2016.51243. , pp. 196-199, 2016.

[28] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *MICCAI,* 2015.

[29] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* p. pages 9190–9200, 2019.