

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

—oo—



HỌC SÂU VÀ ỨNG DỤNG TRONG  
BÀI TOÁN ĐÊM CÂY

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

*Chuyên ngành: TOÁN TIN*

Giảng viên hướng dẫn: TS. LÊ HẢI HÀ

Họ và tên sinh viên: LAI ĐỨC THẮNG

Số hiệu sinh viên: 20163830

Lớp: Toán Tin K61

HÀ NỘI - 2020

# Mục lục

<b>Danh mục từ viết tắt</b>	<b>3</b>
<b>Danh sách hình vẽ</b>	<b>4</b>
<b>Lời cảm ơn</b>	<b>5</b>
<b>Lời mở đầu</b>	<b>6</b>
<b>1 Lý thuyết về hệ thống thông tin địa lý và tiền xử lý ảnh viễn thám</b>	<b>7</b>
1.1 Hệ thống thông tin địa lý . . . . .	7
1.2 Ảnh viễn thám và quá trình tiền xử lý . . . . .	7
<b>2 Lý thuyết học sâu và bài toán nhận diện vật thể</b>	<b>8</b>
2.1 Neural Network - Mạng Neural . . . . .	10
2.2 Mạng Neural tích chập (Convolutional Neural Network - CNN) . . . . .	12
2.3 Bài toán nhận diện vật thể (Object Detection) . . . . .	16
2.3.1 Mô hình Faster R-CNN . . . . .	17
2.3.2 Một số phương pháp đánh giá mô hình nhận diện vật thể . . . . .	23
2.3.3 PyTorch và Faster R-CNN ResNet-50 FPN . . . . .	27
2.3.4 PyTorch . . . . .	28
2.4 Kiến trúc mô hình RetinaNet . . . . .	29
2.4.1 Mạng kim tự tháp đặc trưng Feature Pyramid Network (FPN) .	29
2.4.2 Mạng con phân loại . . . . .	33
2.4.3 Mạng con hồi quy . . . . .	33
2.4.4 Hàm mất mát . . . . .	33

<b>3</b>	<b>Ứng dụng học sâu vào bài toán đếm cây trên ảnh viễn thám</b>	<b>38</b>
3.1	Giới thiệu bài toán . . . . .	38
3.2	Mô hình hoá bài toán và thiết kế dữ liệu luyện . . . . .	39
3.3	Huấn luyện mạng neural . . . . .	44
3.4	Xây dựng chương trình . . . . .	45
<b>4</b>	<b>Cài đặt chương trình và đánh giá kết quả</b>	<b>46</b>
4.1	Môi trường cài đặt chương trình và các yêu cầu liên quan . . . . .	46
4.1.1	Môi trường cài đặt chương trình . . . . .	46
4.1.2	Các yêu cầu liên quan . . . . .	46
4.2	Dữ liệu đầu vào . . . . .	47
4.3	Kết quả mô hình . . . . .	47
4.4	Đánh giá kết quả . . . . .	48
4.5	Định hướng phát triển trong tương lai . . . . .	50
	<b>Tài liệu tham khảo</b>	<b>51</b>

# **Danh mục từ viết tắt**

## **Danh sách hình vẽ**

# Lời cảm ơn

# Lời mở đầu

Việc sản xuất và phân phối loại cây trồng có giá trị kinh tế này đòi hỏi phải có sự giám sát thường xuyên đối với theo chu kỳ và theo mùa vụ. Đây là một nhiệm vụ quan trọng đối với việc quản lý tài nguyên, rừng. Để đáp ứng nhu cầu ngày càng tăng của thế giới và sản lượng nông nghiệp tiêu dùng, năng suất cây trồng phải được ước tính. Việc phát hiện, thu thập thủ công dữ liệu cây trồng để lưu trữ hồ sơ trên một diện tích đất lớn là không khả thi đối với con người, gây tốn kém tiền bạc và dễ bị ảnh hưởng bởi lỗi của con người. Việc phát hiện cây trên một diện tích lớn sẽ giúp doanh nghiệp giảm bớt sự phụ thuộc vào con người và chi phí, dễ dàng quản lý và lên kế hoạch kịp thời cho sản xuất. Trong vài năm trở lại đây, lĩnh vực học sâu và xử lý ảnh đang nhận được nhiều sự quan tâm, áp dụng trong nhiều lĩnh vực khoa học khác nhau. Trong báo cáo đồ án này, em xin giới thiệu một giải pháp phát hiện cây trong các bức ảnh viễn thám dựa trên mô hình học sâu. Nội dung của đồ án gồm có phần mở đầu, 4 chương, phần kết luận, tài liệu tham khảo:

- **Chương 1:**
- **Chương 2:**
- **Chương 3:**

# Chương 1

Lý thuyết về hệ thống thông tin địa lý và tiền xử lý ảnh viễn thám

1.1 Hệ thống thông tin địa lý

1.2 Ảnh viễn thám và quá trình tiền xử lý

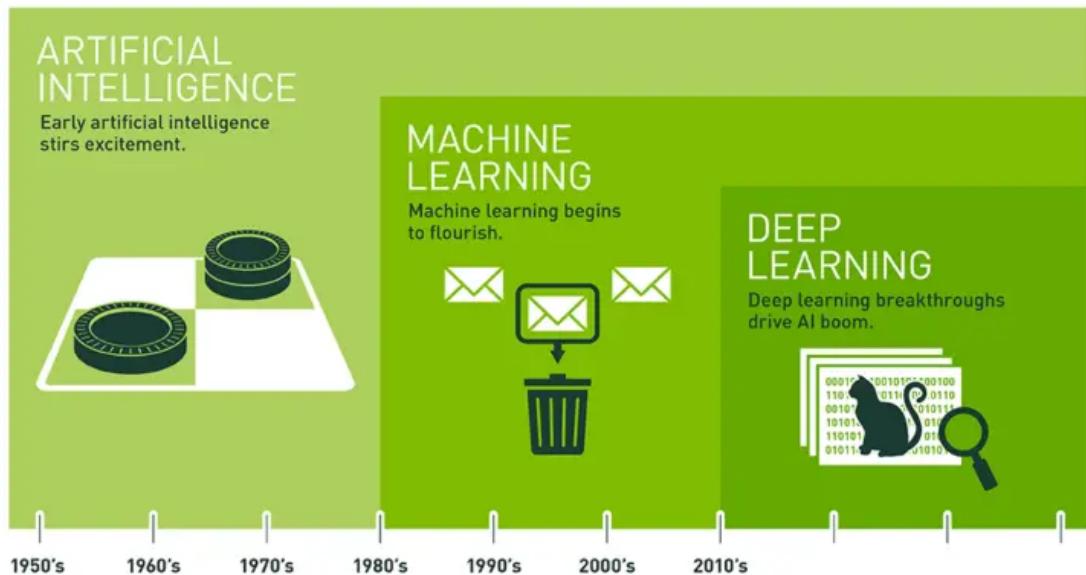
## Chương 2

# Lý thuyết học sâu và bài toán nhận diện vật thể

Deep Learning là một kỹ thuật Machine Learning mà ở đó huấn luyện máy tính giống như cách thức tự nhiên của con người: Học qua các ví dụ. Những năm gần đây, Deep learning đã mang đến nhiều bất ngờ trên quy mô toàn cầu và dẫn đường cho những tiến triển nhanh chóng trong nhiều lĩnh vực khác nhau như thị giác máy tính, xử lý ngôn ngữ tự nhiên (natural language processing), nhận dạng giọng nói tự động (automatic speech recognition), học tăng cường (reinforcement learning), và mô hình hóa thống kê (statistical modeling). Với những tiến bộ này, chúng ta bây giờ có thể xây dựng xe tự lái với mức độ tự động ngày càng cao (nhưng chưa nhiều tới mức như vài công ty đang tuyên bố), xây dựng các hệ thống giúp trả lời thư tự động khi con người ngập trong núi email, hay lập trình phần mềm chơi cờ vây có thể thắng cả nhà vô địch thế giới, một kỷ tích từng được xem là không thể đạt được trong nhiều thập kỷ tới. Những công cụ này đã và đang gây ảnh hưởng rộng rãi tới các ngành công nghiệp và đời sống xã hội, thay đổi cách tạo ra các bộ phim, cách chẩn đoán bệnh và đóng một vài trò ngày càng tăng trong các ngành khoa học cơ bản – từ vật lý thiên văn tới sinh học.

Với Deep Learning, một mô hình máy tính học cách thực hiện một công việc phân loại (classification) trực tiếp từ các hình ảnh, chữ viết (text) hoặc âm thanh. Các mô hình (models) Deep Learning có thể đạt được độ chính xác cao, đôi khi còn hơn cả con người. Các mô hình được huấn luyện bởi việc sử dụng một tập bao gồm bộ dữ liệu

được gán nhãn và các kiến trúc mạng neural gồm nhiều lớp (layer).



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

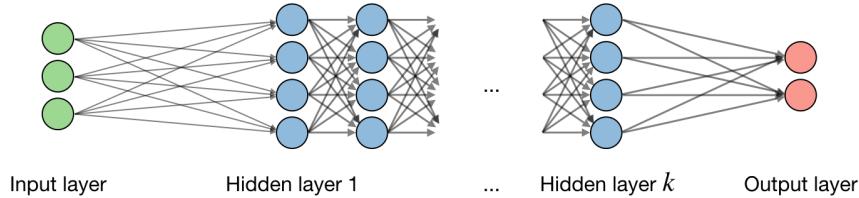
Hình 2.1: Mối quan hệ giữa AI, Machine Learning và Deep Learning. (Nguồn: *What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?*)

**Vậy điều gì mang đến sự thành công của deep learning?** Rất nhiều những ý tưởng cơ bản của deep learning được đặt nền móng từ những năm 80-90 của thế kỷ trước, tuy nhiên deep learning chỉ đột phá trong khoảng từ năm 2012. Vì sao? Có thể kể đến một vài nhân tố dẫn đến sự bùng nổ này:

- Sự ra đời của các bộ dữ liệu lớn được gán nhãn.
- Khả năng tính toán song song tốc độ cao của GPU.
- Sự cải tiến của các kiến trúc: GoogLeNet, VGG, ResNet, ... và các kỹ thuật transfer learning, fine tuning.
- Nhiều thư viện mới hỗ trợ việc huấn luyện deep network với GPU: Theano, Caffe, TensorFlow, PyTorch, Keras,...

## 2.1 Neural Network - Mạng Neural

Tổng quan kiến trúc một mạng Neural như sau



Hình 2.2: Ví dụ một mạng Neural có  $k$  tầng ẩn (Nguồn: CS229)

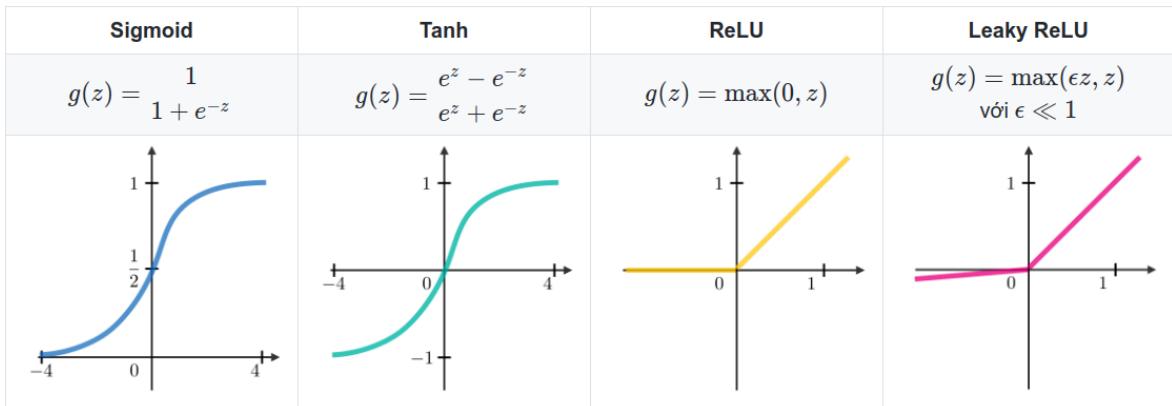
Với  $i$  là lớp thứ  $i$  của mạng,  $j$  là đơn vị ẩn thứ  $j$  của lớp, ta có:

$$z_j^{[i]} = w_j^{[i]T} x + b_j^{[i]}$$

trong đó:  $w$  là weight,  $b$  là bias,  $z$  là đầu ra.

**Hàm kích hoạt (Activation function):** Bản chất của công thức trên là một tổ hợp tuyến tính giữa các giá trị input  $x$  và bộ trọng số  $w$ , do đó, khi áp dụng chúng với các dữ liệu mà có dạng tuyến tính, tức là dữ liệu mà ta có thể kẻ một đường thẳng để phân cách giữa chúng thì công thức tổ hợp trên đã đủ để giúp cho mô hình máy học có thể hoạt động tốt, chúng ta không cần tới hàm kích hoạt (activation function). Nhưng với các dữ liệu không có dạng tuyến tính, ta không thể kẻ một đường thẳng tuyến tính mà phân tách 2 dữ liệu ra được, và câu hỏi đặt ra là làm thế nào với một công thức tổ hợp tuyến tính như ban đầu mà dùng để phân lớp dữ liệu phi tuyến tính được. **Hàm kích hoạt** được tạo ra để làm điều này, hàm kích hoạt đóng vai trò như một người trung gian có nhiệm vụ chuyển đổi, nén hoặc chế biến output  $z$  từ tuyến tính trở thành phi tuyến tính.

**Hàm mất mát (Loss function):** Hàm mất mát trả về một số thực không âm thể hiện sự chênh lệch giữa hai đại lượng:  $y_{pred}$  là giá trị được dự đoán và  $y_{true}$  là giá trị thực. Trong trường hợp lý tưởng,  $y_{pred} = y_{true}$ , hàm mất mát sẽ có giá trị bằng 0. Hàm loss được sử dụng phổ biến trong các mô hình Deep learning hiện nay là Cross-entropy cùng các biến thể cải tiến của nó (Weighted cross entropy, Focal loss...). Cross-entropy



Hình 2.3: Một số hàm kích hoạt thường dùng (Nguồn: CS229)

loss  $L(z, y)$  được định nghĩa như sau:

$$L(z, y) = -[y \log z + (1 - y) \log(1 - z)]$$

**Optimizer và Learning rate:** Sau khi tính giá trị hàm loss, việc cần làm là tối ưu (cực tiểu hóa) hàm loss và update bộ trọng số  $\{w\}$  mới. Learning rate, thường được ký hiệu là  $\alpha$  hoặc  $\eta$ , thể hiện cho tốc độ học hay tốc độ update trọng số. Learning rate có thể là cố định hoặc được thay đổi tùy biến trong quá trình học.

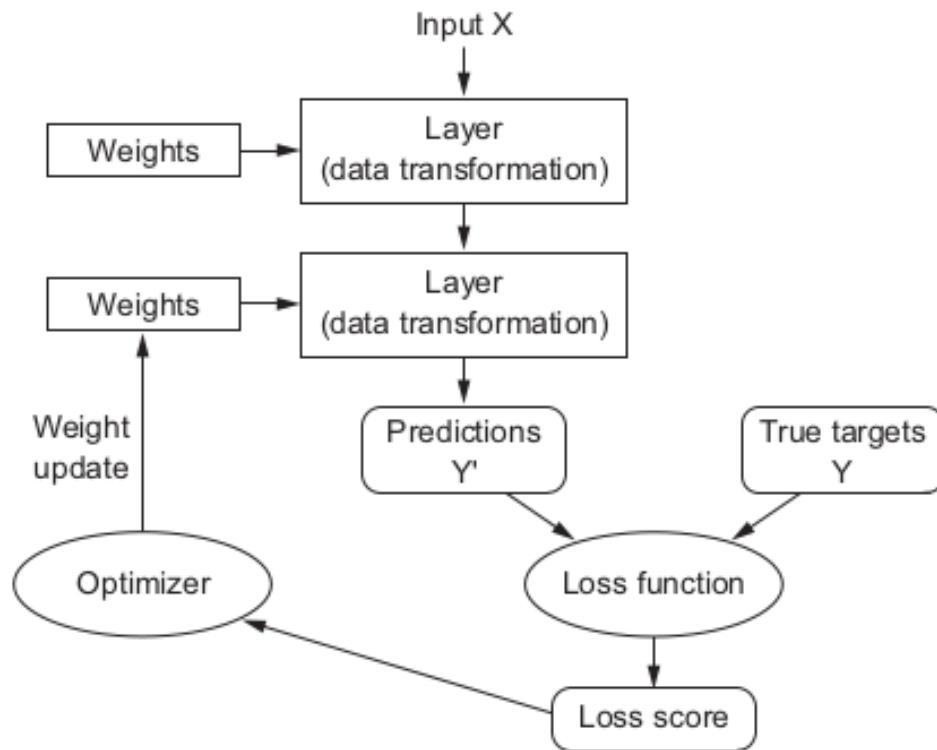
**Lan truyền ngược (Backpropagation):** Lan truyền ngược là phương thức dùng để cập nhật trọng số trong mạng neural bằng cách tính toán đầu ra thực sự và đầu ra mong muốn. Đạo hàm theo trọng số  $w$  được tính bằng cách sử dụng quy tắc chuỗi (chain rule) dưới đây:

$$\frac{\partial L(z, y)}{\partial w} = \frac{\partial L(z, y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w}$$

Như kết quả, trọng số được cập nhật như sau:

$$w := w - \eta \frac{\partial L(z, y)}{\partial w}$$

Tổng kết lại, ta có sơ đồ quá trình học của một mạng neural cơ bản như sau



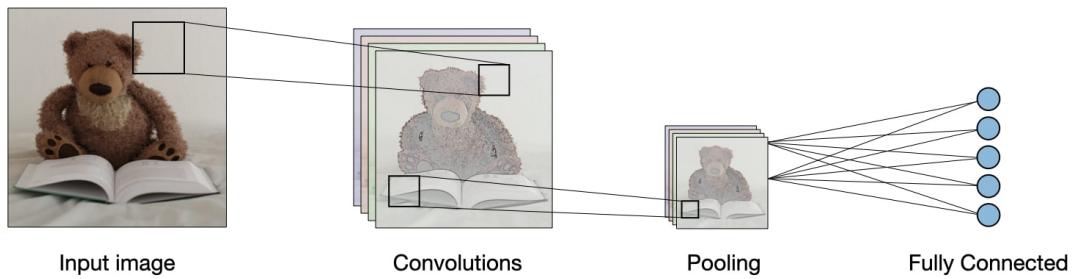
Hình 2.4: Mối quan hệ giữa network, layers, loss function và optimizer (Nguồn: Deep Learning with Python - Francois Chollet)

## 2.2 Mạng Neural tích chập (Convolutional Neural Network - CNN)

Mạng Neural tích chập là một trong những mô hình Deep learning phổ biến nhất và có ảnh hưởng nhiều nhất trong lĩnh vực Computer Vision. CNNs được dùng trong nhiều bài toán như nhận dạng ảnh, phân tích video, ảnh MRI, hoặc có thể cho cả các bài của lĩnh vực xử lý ngôn ngữ tự nhiên, và hầu hết đều giải quyết tốt các bài toán này.

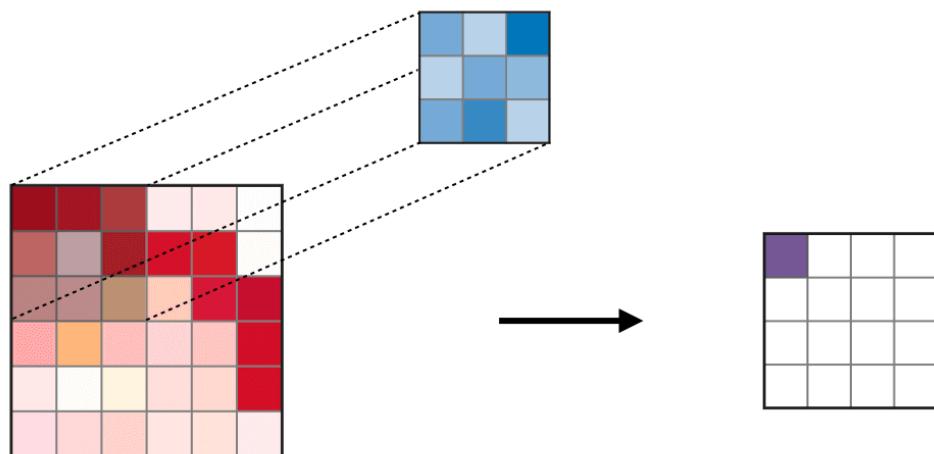
Mạng neural tích chập, còn được biết đến với tên CNNs, là một dạng mạng neural được cấu thành bởi các layer sau:

**Lớp tích chập (Convolution layer):** Tầng tích chập (CONV) sử dụng các bộ lọc (filters) để thực hiện phép tích chập khi đưa chúng đi qua input  $I$  theo các chiều của nó. Các *hyperparameters* của filter bao gồm kích thước  $F$ , độ trượt (stride)  $S$ . Kết quả



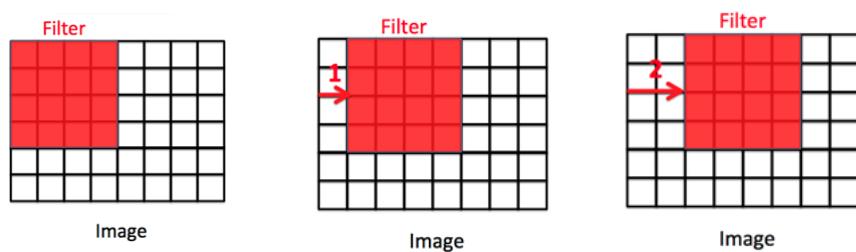
Hình 2.5: Ví dụ về một CNN (Nguồn: CS230)

đầu ra của lớp này được gọi là feature map.



Hình 2.6: Mô tả hoạt động của CONV (Nguồn: CS230)

**Stride** là số lượng pixel dịch chuyển trên ma trận đầu vào hay Stride dùng để dịch chuyển filter theo mỗi bước xác định.



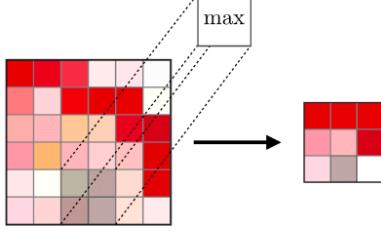
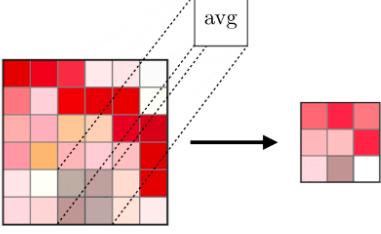
Ví dụ về stride = 1 và stride bằng 2

**Padding:** Khi áp dụng phép CONV thì ma trận đầu vào sẽ có nhỏ dần đi, do đó số layer của mô hình CNN sẽ bị giới hạn, và không thể xây dựng mô hình mong muốn.

Để giải quyết tình trạng này, ta cần "bọc" bên ngoài ma trận đầu vào để đảm bảo kích thước đầu ra sau mỗi tầng convolution là không đổi. Do đó có thể xây dựng được mô hình với số tầng convolution lớn tùy ý. Một cách đơn giản và phổ biến nhất để padding là sử dụng hàng số 0, ngoài ra có một số phương pháp khác như reflection padding hay là symmetric padding.

**Lớp Pooling:** Pooling layer thường được dùng giữa các convolutional layer, để giảm kích thước dữ liệu nhưng vẫn giữ được các thuộc tính quan trọng. Kích thước dữ liệu giảm giúp giảm việc tính toán trong model.

**Spatial pooling** được gọi là lấy mẫu con làm giảm chiều của mỗi map nhưng vẫn giữ được thông tin quan trọng. Spatial pooling có thể có nhiều loại khác nhau như Max Pooling và Average Pooling.

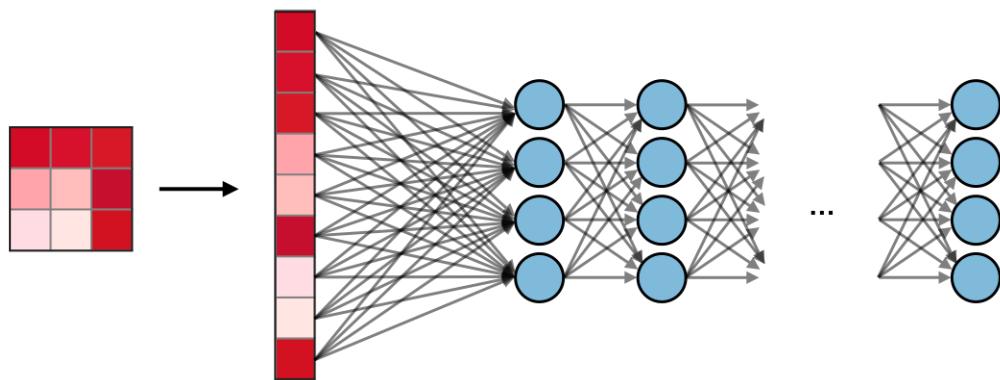
Kiểu	Max pooling	Average pooling
<b>Chức năng</b>	Từng phép pooling chọn giá trị lớn nhất trong khu vực mà nó đang được áp dụng	Từng phép pooling tính trung bình các giá trị trong khu vực mà nó đang được áp dụng
<b>Mình họa</b>		
<b>Nhận xét</b>	<ul style="list-style-type: none"> <li>Bảo toàn các đặc trưng đã phát hiện</li> <li>Được sử dụng thường xuyên</li> </ul>	<ul style="list-style-type: none"> <li>Giảm kích thước feature map</li> <li>Được sử dụng trong mạng LeNet</li> </ul>

Hình 2.7: Hai kiểu pooling phổ biến (Nguồn: CS230)

**Fully Connected (FC):** Lớp Fully Connected nhận đầu vào là các dữ liệu đã được làm phẳng, mà mỗi đầu vào đó được kết nối đến tất cả neuron. Trong mô hình mạng CNNs, các lớp FC thường được tìm thấy ở cuối mạng và được dùng để tối ưu hóa mục tiêu của mạng ví dụ như độ chính xác của lớp.

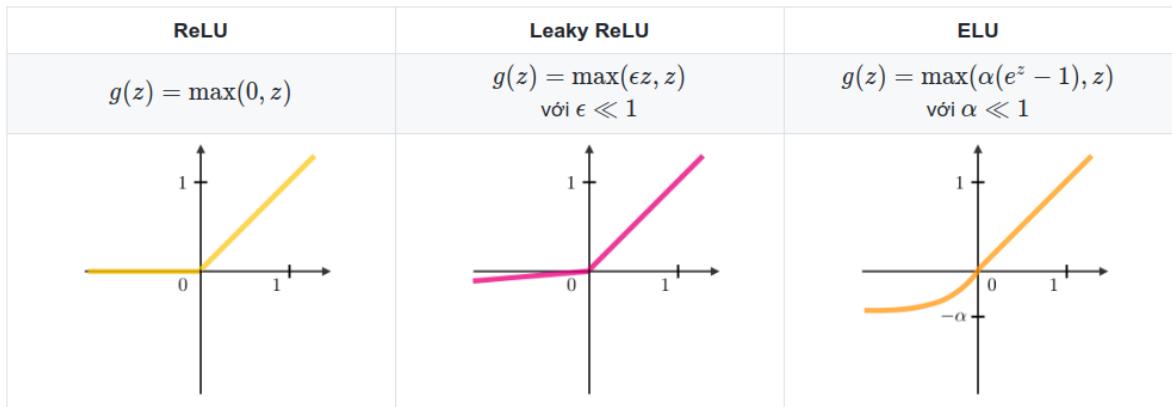
**Các hàm kích hoạt thường gặp:**

- **Rectified Linear Unit (ReLU):** ReLU là một hàm kích hoạt  $g$  được sử dụng trên tất cả các thành phần. Mục đích của nó là tăng tính phi tuyến tính cho



Hình 2.8: Fully Connected layer (Nguồn: CS230)

mạng. Những biến thể khác của ReLU được tổng hợp ở bảng dưới



Hình 2.9: ReLU và biến thể (Nguồn: CS230)

- **Softmax:** Bước softmax có thể được coi là một hàm logistic tổng quát lấy đầu vào là một vector chứa các giá trị  $x \in \mathbb{R}^n$  và cho ra là một vector gồm các xác suất  $p \in \mathbb{R}^n$  thông qua một hàm softmax ở layer cuối.

$$p = (p_1, \dots, p_n)^T \text{ trong đó } p = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}$$

## 2.3 Bài toán nhận diện vật thể (Object Detection)

Các hình ảnh trong cuộc sống bình thường thì không chỉ chứa 1 đối tượng mà thường bao gồm rất nhiều các đối tượng. Ta quan tâm đến vị trí của từng đối tượng trong ảnh. Bài toán như vậy được gọi là: object detection.



Hình 2.10: Ví dụ về đầu ra của bài toán nhận diện vật thể

Bài toán object detection có input là ảnh màu và output là vị trí của các đối tượng trong ảnh. Ta thấy nó bao gồm 2 bài toán nhỏ:

- Xác định các bounding box (hình chữ nhật) quanh đối tượng.
- Với mỗi bounding box thì cần phân loại xem đây là đối tượng gì (người, mèo, ô tô,...) với bao nhiêu phần trăm chắc chắn.

Object Detection là 1 bài toán đã đạt được rất nhiều các thành tựu trong những năm gần đây, cả phần ứng dụng và mô hình thuật toán. Diễn hình là các phương pháp Object Detection sử dụng Deep Learning đã đạt được các bước cải thiện vượt trội so với các phương pháp xử lý ảnh thông thường khác.

Có hai loại bài toán Object detection: two-stage object detection và one-stage object detection.

**Two-stage object detection:** Diễn hình họ các thuật toán R-CNN. Việc gọi là two-stage là do cách model xử lý để lấy ra được các vùng có khả năng chứa vật thể từ bức ảnh. Ví dụ, với Faster-RCNN thì trong stage-1, ảnh sẽ được đưa ra 1 sub-network gọi là RPN (Region Proposal Network) với nhiệm vụ extract các vùng trên ảnh có khả năng chứa đối tượng dựa vào các anchor. Sau khi đã thu được các vùng đặc trưng từ RPN, model Faster-RCNN sẽ thực hiện tiếp việc phân loại đối tượng và xác định vị trí nhờ vào việc chia làm 2 nhánh tại phần cuối của mô hình (Object classification & Bounding box regression).

**One-stage Object Detection:** Các thuật toán diễn hình như: SSD, YOLO, RetinaNet. Gọi là one-stage vì trong việc thiết kế model hoàn toàn không có phần trích chọn các vùng đặc trưng (các vùng có khả năng chứa đối tượng) như RPN của Faster-RCNN. Các mô hình one-stage object detection coi phần việc phát hiện đối tượng (object localization) như một bài toán regression (với 4 tọa độ offset, ví dụ x, y, w, h) và cũng dựa trên các box được định nghĩa sẵn gọi là anchor để làm việc đó. Các mô hình dạng này thường nhanh hơn tuy nhiên "độ chính xác" của model thường kém hơn so với two-stage object detection. Tuy nhiên, một số mô hình one-stage vẫn tỏ ra vượt trội hơn một chút so với two-stage như Retina-Net với việc việc thiết kế mạng theo FPN (Feature Pyramid Network) và Focal Loss.

### 2.3.1 Mô hình Faster R-CNN

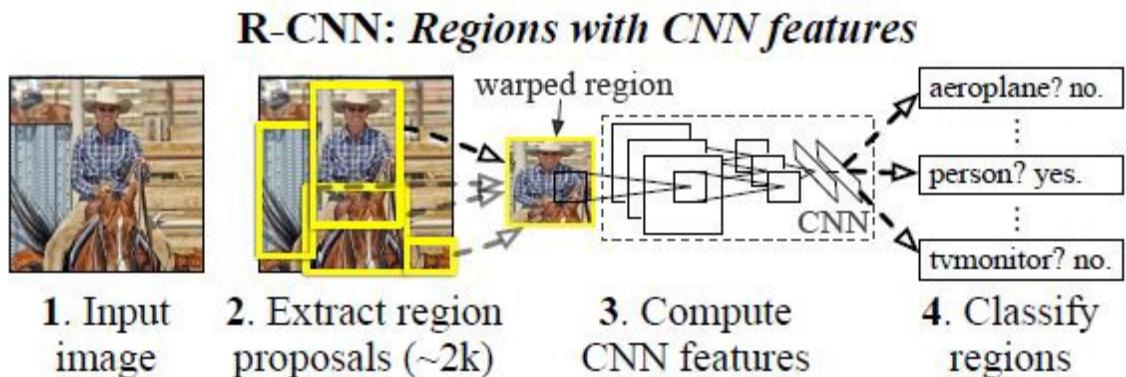
#### Regions with CNN features (R-CNN)

R-CNN được giới thiệu lần đầu vào 2014 bởi Ross Girshick và các cộng sự ở UC Berkeley một trong những trung tâm nghiên cứu AI hàng đầu thế giới trong bài báo *Rich feature hierarchies for accurate object detection and semantic segmentation*. RCNN có thể là xem một trong những ứng dụng nền móng đầu tiên của CNN đối với bài toán định vị, phát hiện và phân đoạn đối tượng.

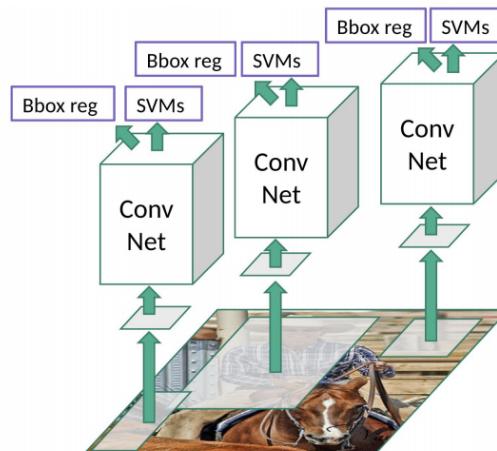
Kiến trúc của R-CNN gồm 3 thành phần đó là:

- Vùng đề xuất hình ảnh (Region proposal): Có tác dụng tạo và trích xuất các vùng đề xuất chứa vật thể được bao bởi các bounding box.

- Trích lọc đặc trưng (Feature Extractor): Trích xuất các đặc trưng giúp nhận diện hình ảnh từ các region proposal thông qua các CNN.
- Phân loại (classifier): Dựa vào input là các features ở phần trước để phân loại hình ảnh chứa trong region proposal về đúng nhãn.



Hình 2.11: Sơ đồ xử lý trong mô hình mạng R-CNN (Nguồn: Medium)



Hình 2.12: R-CNN (Nguồn: Medium)

Để vượt qua vấn đề chọn một số lượng lớn các region, Ross Girshick đề xuất một phương pháp gọi là selective search để chỉ trích xuất 2000 regions từ hình ảnh và gọi chúng là vùng đề xuất (region proposals), tức vùng có khả năng chứa đối tượng. Do đó, thay vì cố gắng phân loại một số lượng lớn các region, ta chỉ cần làm việc với 2000 regions. 2000 proposal regions này được uốn cong thành một hình vuông và được đưa

vào một mạng CNN tạo ra một vector đặc trưng 4096 chiều làm đầu ra. CNN hoạt động như một công cụ trích xuất đặc trưng và dense layer đầu ra bao gồm các đặc trưng được trích xuất từ các ảnh và các đặc trưng được đưa vào một SVM để phân loại sự hiện diện của đối tượng trong proposal region được lấy ra đó. Ngoài việc dự đoán sự hiện diện của một đối tượng trong các proposal region, thuật toán cũng dự đoán bốn giá trị là độ lệch (offset values) để tăng độ chính xác của bounding box, Ví dụ với một region proposal cho trước, thuật toán sẽ dự đoán sự hiện diện của người nhưng khuôn mặt người đó trong region proposal đó có thể bị cắt đi một nửa. Do đó, các giá trị bù giúp điều chỉnh bounding box của region proposal.

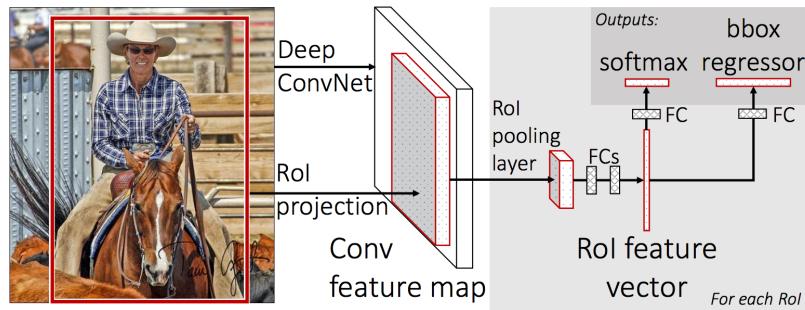
#### Nhược điểm của R-CNN:

- Cần tốn lượng lớn thời gian để luyện mạng vì phải phân loại 2000 region proposals mỗi ảnh.
- Không thể chạy với thời gian thực do nó tốn thời gian 47 giây cho mỗi ảnh
- Thuật toán selective search là một thuật toán cố định. Do đó, không có việc học nào đang diễn ra ở giai đoạn đó, Điều này có thể dẫn đến việc tạo ra các region proposal tồi.

## Fast R-CNN

Dựa trên thành công của R-CNN, Ross Girshick (lúc này đã chuyển sang Microsoft Research) đề xuất một mở rộng để giải quyết vấn đề của R-CNN trong một bài báo vào năm 2015 với tiêu đề rất ngắn gọn Fast R-CNN.

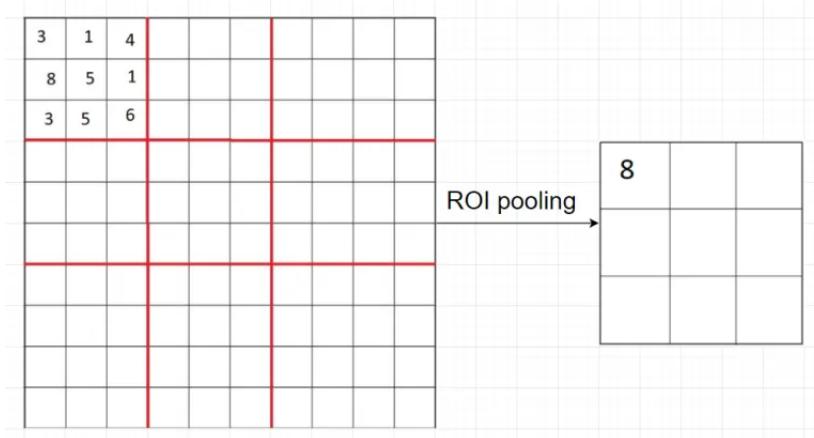
Tương tự như R-CNN thì Fast R-CNN vẫn dùng selective search để lấy ra các region proposal. Tuy nhiên là nó không tách 2000 region proposal ra khỏi ảnh và thực hiện bài toán image classification cho mỗi ảnh. Fast R-CNN cho cả bức ảnh vào ConvNet (một vài convolutional layer + max pooling layer) để tạo ra convolutional feature map. Sau đó các vùng region proposal được lấy ra tương ứng từ convolutional feature map. Tiếp đó được Flatten và thêm 2 Fully connected layer (FCs) để dự đoán lớp của region proposal và offset values của bounding box.



Hình 2.13: Fast R-CNN (Nguồn: Medium)

Tuy nhiên là kích thước của các region proposal khác nhau nên khi Flatten sẽ ra các vector có kích thước khác nhau nên không thể áp dụng neural network được. Với R-CNN, nó đã resize các region proposal về cùng kích thước trước khi dùng transfer learning. Tuy nhiên ở feature map ta không thể resize được, nên ta phải có cách gì đó để chuyển các region proposal trong feature map về cùng kích thước nên **Region of Interest (ROI) pooling** ra đời.

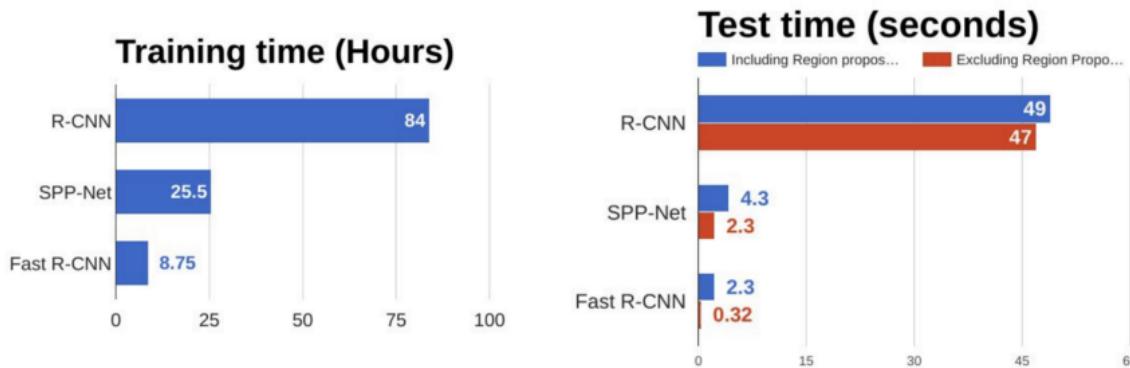
**Region of Interest (ROI) pooling:** ROI pooling là một dạng của pooling layer. Điểm khác so với max pooling hay average pooling là bất kể kích thước của tensor input, ROI pooling luôn cho ra output có kích thước cố định được định nghĩa trước.



Hình 2.14: ROI pooling (Nguồn: ntuan8.com)

Fast R-CNN khác với R-CNN là nó thực hiện feature map với cả ảnh sau đó với lấy các region proposal ra từ feature map, còn R-CNN thực hiện tách các region proposal ra rồi mới thực hiện CNN trên từng region proposal. Do đó Fast R-CNN nhanh hơn

đáng kể nhờ tối ưu việc tính toán bằng Vectorization.



Hình 2.15: So sánh một số thuật toán object detection (Nguồn: [towardsdatascience.com](https://towardsdatascience.com/))

Tuy nhiên nhìn hình trên ở phần test time với mục Fast R-CNN thì thời gian tính region proposal rất lâu và làm chậm thuật toán. Do đó cần thay thế thuật toán selective search và việc dùng Deep learning để tạo ra region proposal được thực hiện với mô hình Faster R-CNN.

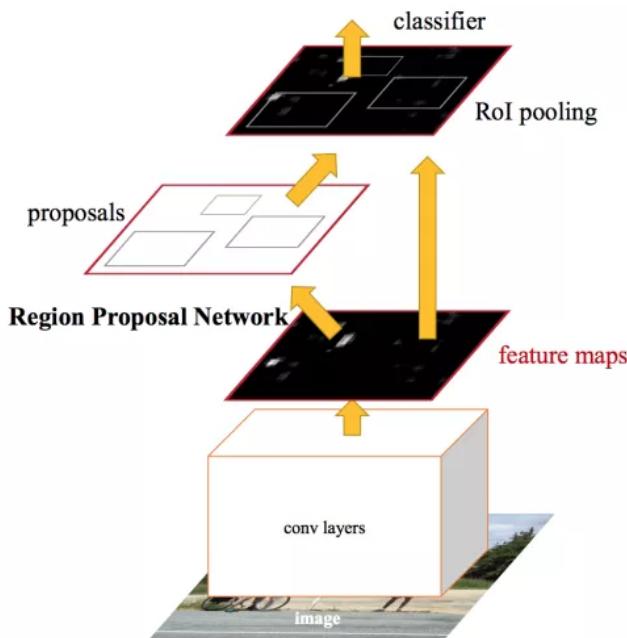
## Faster R-CNN

Faster R-CNN không dùng thuật toán selective search để lấy ra các region proposal, mà nó thêm một mạng CNN mới gọi là Region Proposal Network (RPN) để tìm các region proposal.

Một Region Proposal Network nhận đầu vào là ảnh với kích thước bất kỳ và cho đầu ra là region proposal (tập vị trí của các hình chữ nhật có thể chứa vật thể), cùng với xác suất chứa vật thể của hình chữ nhật tương ứng.

**Region Proposal Network (RPN):** Quy trình tính toán của RPN được mô tả chi tiết dưới đây:

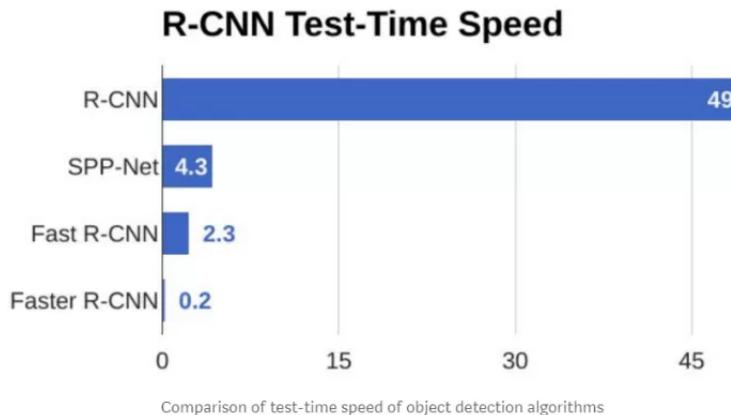
1. Dùng một lớp tích chập  $3 \times 3$  với padding bằng 1 để biến đổi đầu ra của CNN và đặt số kênh đầu ra bằng  $c$ . Bằng cách này phần tử trong feature map mà CNN trích xuất ra từ bức ảnh là một đặc trưng mới có độ dài bằng  $c$ .



Hình 2.16: Mô hình Faster R-CNN (Nguồn: *arXiv:1506.01497*)

2. Lấy mỗi phần tử trong feature map làm tâm để tạo ra nhiều anchor box có kích thước và tỷ lệ khác nhau, sau đó gán nhãn cho chúng.
3. Lấy những đặc trưng của các phần tử có độ dài  $c$  ở tâm của anchor box để phân loại nhị phân (là vật thể hay là nền) và dự đoán bounding box tương ứng cho các anchor box.
4. Sau đó, sử dụng non-maximum suppression để loại bỏ các bounding box có kết quả giống nhau của hạng mục “vật thể”. Cuối cùng, ta xuất ra các bounding box dự đoán là các proposal region rồi đưa vào lớp RoI pooling.

Vì là một phần của mô hình Faster R-CNN, nên RPN được huấn luyện cùng với phần còn lại trong mô hình. Ngoài ra, trong đối tượng Faster R-CNN còn chứa các hàm dự đoán hạng mục và bounding box trong bài toán phát hiện vật thể, cũng như các hàm dự đoán hạng mục nhị phân và bounding box cho các anchor box trong RPN. Sau cùng, RPN có thể học được cách sinh ra những proposal region có chất lượng cao, giảm đi số lượng proposal region trong khi vẫn giữ được độ chính xác khi phát hiện vật thể.



Hình 2.17: Faster R-CNN nhanh hơn hẳn các dòng R-CNN trước đó, vì vậy có thể dùng cho real-time objec detection (Nguồn: [nttuan8.com](http://nttuan8.com))

### 2.3.2 Một số phương pháp đánh giá mô hình nhận diện vật thể

#### Intersection over Union (IoU)

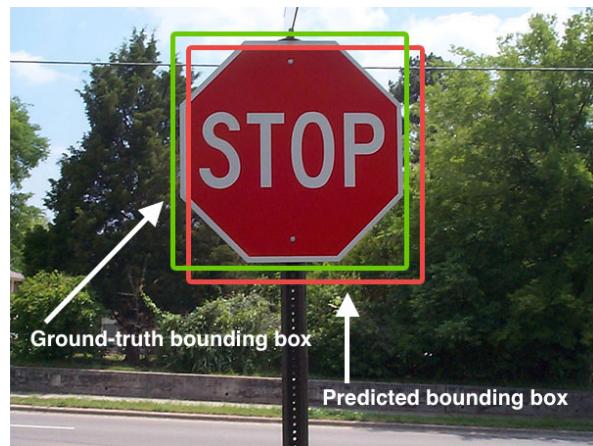
Intersection over Union là chỉ số đánh giá được sử dụng để đo độ chính xác của Object detector trên tập dữ liệu cụ thể. IoU đơn giản chỉ là một chỉ số đánh giá. Mọi thuật toán có khả năng predict ra các bounding box làm output đều có thể được đánh giá thông qua IoU.

Để áp dụng được IoU để đánh giá một mô hình nhận diện vật thể bất kì ta cần:

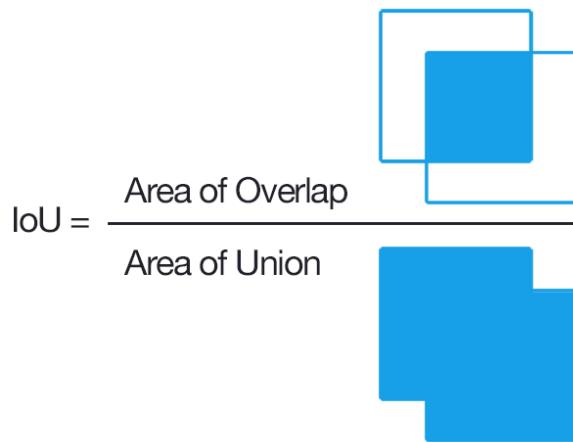
- Những ground-truth bounding box (bounding box đúng của đối tượng, ví dụ như bounding box của đối tượng được khoanh vùng và gán nhãn bằng tay sử dụng trong tập test.)
- Những bounding box dự đoán được model sinh ra.

#### Average Precision

AP (Average Precision) là một metric phổ biến trong việc đánh giá độ chính xác của các mô hình nhận diện vật thể như aster R-CNN, SSD,...



Hình 2.18: Một ví dụ về phát hiện biển báo Stop từ hình ảnh. (Nguồn: [pyimagesearch.com](http://pyimagesearch.com))



Hình 2.19: Tính toán Intersection over Union. (Nguồn: [pyimagesearch.com](http://pyimagesearch.com))

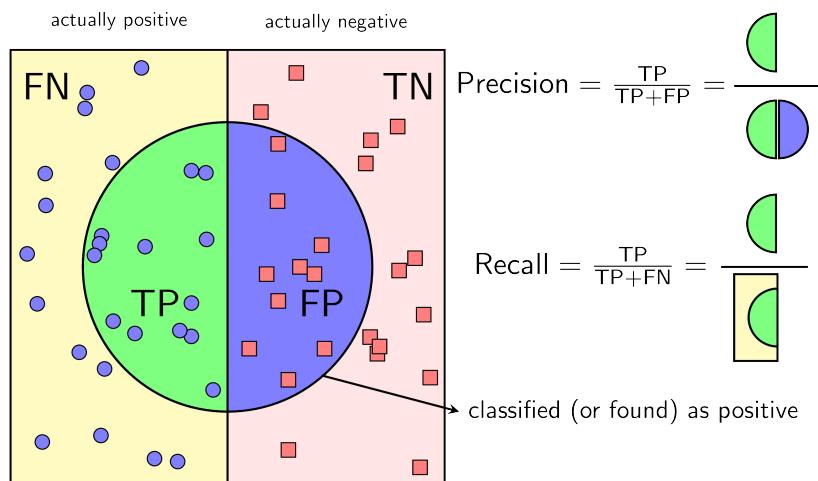


Hình 2.20: Một ví dụ về tính toán IoU cho những bounding box khác nhau. (Nguồn: [pyimagesearch.com](http://pyimagesearch.com))

## Precision – Recall

Với bài toán phân loại mà tập dữ liệu của các lớp là chênh lệch nhau rất nhiều, có một phép đo hiệu quả thường được sử dụng là Precision-Recall.

Trước hết xét bài toán phân loại nhị phân. Ta cũng coi một trong hai lớp là positive, lớp còn lại là negative.



Hình 2.21: Cách tính Precision và Recall. (Nguồn: *Machine Learning cơ bản*)

**Precision** cao đồng nghĩa với việc độ chính xác của các điểm tìm được là cao. **Recall** cao đồng nghĩa với việc True Positive Rate cao, tức tỉ lệ bỏ sót các điểm thực sự positive là thấp.

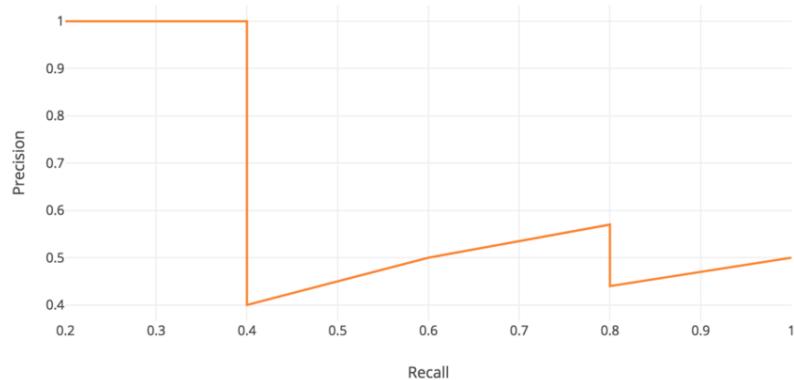
### Precision-Recall curve và Average precision

Ta có thể đánh giá mô hình dựa trên việc thay đổi một ngưỡng và quan sát giá trị của Precision và Recall.

**Average Precision:** Ta xét một ví dụ, với bộ dữ liệu có 5 quả táo, mô hình lần lượt đưa ra 10 dự đoán, ta chọn ngưỡng cho dự đoán đúng là  $\text{IoU} > 0.5$ , dự đoán được xếp giảm dần theo sự 'tự tin' của dự đoán. Số liệu được thể hiện trong bảng sau:

Rank	Correct?	Precision	Recall
1	True	1.0	0.2
2	True	1.0	0.4
3	False	0.67	0.4
4	False	0.5	0.4
5	False	0.4	0.4
6	True	0.5	0.6
7	True	0.57	0.8
8	False	0.5	0.8
9	False	0.44	0.8
10	True	0.5	1.0

Biểu diễn các điểm precision-recall ta được một đường zig-zag sau



Hình 2.22: Precision-recall curve (Nguồn: Medium)

Dịnh nghĩa tổng quát cho Average Precision (AP) là diện tích phía dưới precision-recall curve

$$AP = \int_0^1 p(r)dr$$

Precision và recall luôn nằm trong đoạn [0; 1] do đó AP cũng nằm trong đoạn [0; 1].

Mean average precision (mAP) là trung bình của AP. Trong một số trường hợp, ta tính AP cho mỗi class và lấy trung bình của chúng, một số khác thì lại giống nhau. Ví dụ theo COCO, không có sự khác biệt giữa AP và mAP.

<b>Average Precision (AP) :</b>	
AP	% AP at IoU=.50:.05:.95 ( <b>primary challenge metric</b> )
AP <sub>IoU=.50</sub>	% AP at IoU=.50 (PASCAL VOC metric)
AP <sub>IoU=.75</sub>	% AP at IoU=.75 (strict metric)
<b>AP Across Scales:</b>	
AP <sub>small</sub>	% AP for small objects: area < 32 <sup>2</sup>
AP <sub>medium</sub>	% AP for medium objects: 32 <sup>2</sup> < area < 96 <sup>2</sup>
AP <sub>large</sub>	% AP for large objects: area > 96 <sup>2</sup>
<b>Average Recall (AR) :</b>	
AR <sub>max=1</sub>	% AR given 1 detection per image
AR <sub>max=10</sub>	% AR given 10 detections per image
AR <sub>max=100</sub>	% AR given 100 detections per image
<b>AR Across Scales:</b>	
AR <sub>small</sub>	% AR for small objects: area < 32 <sup>2</sup>
AR <sub>medium</sub>	% AR for medium objects: 32 <sup>2</sup> < area < 96 <sup>2</sup>
AR <sub>large</sub>	% AR for large objects: area > 96 <sup>2</sup>

Hình 2.23: Một số metric được sử dụng để đánh giá kết quả trên bộ dữ liệu COCO (Nguồn: [cocodataset.org](http://cocodataset.org))

### 2.3.3 PyTorch và Faster R-CNN ResNet-50 FPN

#### Sơ lược về Transfer Learning

Những năm gần đây, Deep Learning phát triển cực nhanh dựa trên lượng dữ liệu training khổng lồ và khả năng tính toán ngày càng được cải tiến của các máy tính. Các kết quả cho bài toán phân loại ảnh ngày càng được nâng cao. Bộ cơ sở dữ liệu thường được dùng nhất là ImageNet với 1.2M ảnh cho 1000 classes khác nhau. Rất nhiều các mô hình Deep Learning đã giành chiến thắng trong các cuộc thi ILSVRC (ImageNet Large Scale Visual Recognition Challenge). Có thể kể ra một vài: AlexNet, ZFNet, GoogLeNet, ResNet, VGG.

Nhìn chung, các mô hình này đều bao gồm rất nhiều layers. Các layers phía trước thường là các Convolutional layers kết hợp với các nonlinear activation functions và pooling layers (và được gọi chung là ConvNet). Layer cuối cùng là một Fully Connected Layer và thường là một Softmax Regression (Xem Hình 1). Số lượng units ở layer cuối cùng bằng với số lượng classes (với ImageNet là 1000). Vì vậy output ở layer gần cuối cùng (second to last layer) có thể được coi là feature vectors và Softmax Regression chính là Classifier được sử dụng.

Chính nhờ việc features và classifier được trained cùng nhau qua deep networks khiến

cho các mô hình này đạt kết quả tốt. Tuy nhiên, những mô hình này đều là các Deep Networks với rất nhiều layers. Việc training dựa trên 1.2M bức ảnh của ImageNet cũng tốn rất nhiều thời gian (2-3 tuần).

Với các bài toán dựa trên tập dữ liệu khác, rất ít khi người ta xây dựng và train lại toàn bộ Network từ đầu, bởi vì có rất ít các cơ sở dữ liệu có kích thước lớn. Thay vào đó, phương pháp thường được dùng là sử dụng các mô hình (nêu phía trên) đã được trained từ trước, và sử dụng một vài kỹ thuật khác để giải quyết bài toán. Phương pháp sử dụng các mô hình có sẵn như thế này được gọi là Transfer Learning.

Có 2 loại transfer learning:

- **Feature extractor:** Sau khi lấy ra các đặc điểm của ảnh bằng việc sử dụng ConvNet của pre-trained model, thì ta sẽ dùng linear classifier (linear SVM, softmax classifier,...) để phân loại ảnh.
- **Fine tuning:** Sau khi lấy ra các đặc điểm của ảnh bằng việc sử dụng ConvNet của pre-trained model, thì ta sẽ coi đây là input của 1 CNN mới bằng cách thêm các ConvNet và Fully Connected layer.

### 2.3.4 PyTorch

PyTorch là một thư viện machine learning mã nguồn mở dựa trên Torch, được sử dụng cho lĩnh vực Thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural language processing), được phát triển bởi Phòng nghiên cứu AI của Facebook (FAIR).

Pytorch tập trung vào 2 khả năng chính:

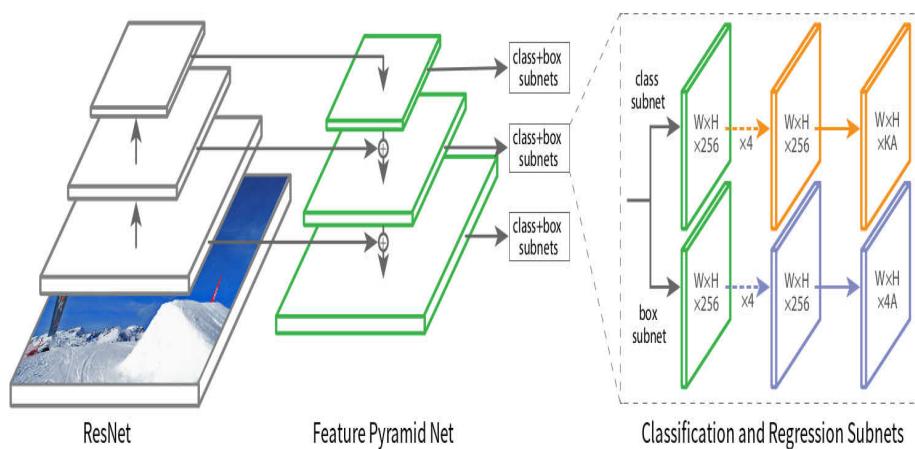
- Một sự thay thế cho bộ thư viện numpy để tận dụng sức mạnh tính toán của GPU.
- Một platform Deep learning phục vụ trong nghiên cứu, mang lại sự linh hoạt và tốc độ.

Cấu trúc dữ liệu cốt lõi được sử dụng trong PyTorch là **Tensor**.

## 2.4 Kiến trúc mô hình RetinaNet

RetinaNet là một mạng tổng hợp bao gồm:

- Mạng kim tự tháp đặc trưng-Feature Pyramid Networks (FPN) được xây dựng dựa trên mạng ResNet; chịu trách nhiệm tính toán, trích xuất các đặc trưng trong bức ảnh
- Mạng con chịu trách nhiệm thực hiện phân loại đối tượng bằng cách sử dụng đầu ra của mạng kim tự tháp đặc trưng FPN
- Mạng con chịu trách nhiệm điều chỉnh các kích thước phát hiện vật thể từ đầu ra của mạng kim tự tháp đặc trưng FPN

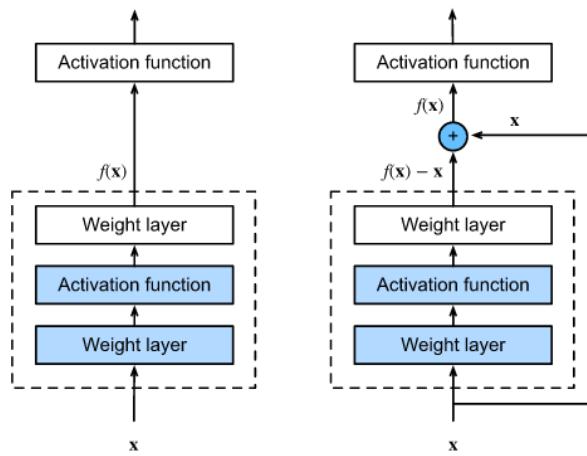


Hình 2.24: Kiến trúc mô hình RetinaNet.

### 2.4.1 Mạng kim tự tháp đặc trưng Feature Pyramid Network (FPN)

**Residual Network (ResNet):** Một vấn đề phổ biến của Deep learning là Vanishing Gradients, tức là theo công thức tính đạo hàm bằng chain rule (trong lan truyền ngược, back propagation), nếu gradient của các lớp sau nhỏ thì gradient ở các lớp đầu sẽ gần bằng 0. Vì thế mà parameter của các lớp trước sẽ không được cập nhật nên các

parameter này không đóng góp được gì trong việc đưa ra output. Vậy nên dù ta dùng nhiều lớp, thực chất số lớp hữu ích chỉ là một vài lớp cuối, điều đó làm giảm sự hiệu quả của mạng neural. ResNet ra đời để giải quyết vấn đề đó, 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.



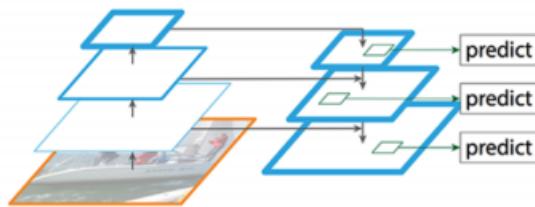
Hình 2.25: Khối thường (trái) và khối residual (phải)

ResNet-50 là một mạng bao gồm 50 lớp residual.

### Mạng kim tự tháp đặc trưng-Feature Pyramid Networks (FPN)

Phát hiện các đối tượng có kích thước nhỏ là một vấn đề vô cùng quan trọng để nâng cao độ chính xác. Và FPN là mô hình mạng được thiết kế ra dựa trên khái niệm kim tự tháp (pyramid) để giải quyết vấn đề này.

Mô hình FPN giúp nhận diện được cả những đối tượng có kích thước to và nhỏ

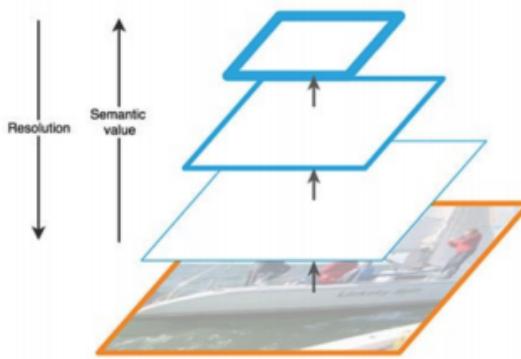


Hình 2.26: Feature Pyramid Network (FPN)

nhờ sự kết hợp của tất cả các lớp kim tự tháp. Mô hình được xây dựng bằng cách thực hiện quá trình bottom-up kết hợp với top-down để dò tìm đối tượng (trong khi đó, các

thuật toán khác chỉ thường sử dụng bottom-up). Trong đó bottom-up sử dụng mạng tích chập thông thường dùng để trích xuất các đặc trưng, khi chúng ta ở bottom và đi lên (up) độ phân giải sẽ giảm, nhưng giá trị cấu trúc cấp cao (high-level semantic) sẽ tăng lên.

Một vài các mô hình khác trong phát hiện đối tượng (ví dụ: SSD,...) quyết định

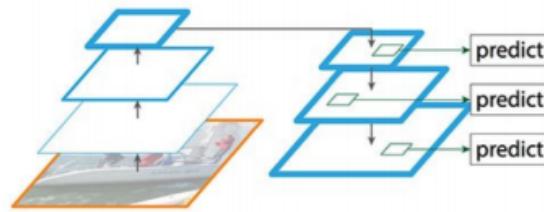


Hình 2.27: Trích xuất đặc trưng trong mô hình FPN

dựa vào nhiều feature map. Nhưng tầng ở dưới không được sử dụng để nhận dạng đối tượng. Vì những tầng này có độ phân giải cao nhưng giá trị cấu trúc cấp cao của chúng lại không đủ cao nên những nhà nghiên cứu bỏ chúng đi để tăng tốc độ xử lý. Các nhà nghiên cứu biện minh rằng các tầng ở dưới chưa đủ mức ý nghĩa cần thiết để nâng cao độ chính xác, thêm các tầng đó vào sẽ không nâng độ chính xác thêm bao nhiêu và họ bỏ chúng đi để có tốc độ tốt hơn. Cho nên, những mô hình phát hiện đối tượng này chỉ sử dụng các tầng ở lớp trên, và do đó sẽ không nhận dạng được các đối tượng có kích thước nhỏ.

Trong khi đó, FPN xây dựng thêm mô hình top-down, nhằm mục đích xây dựng các tầng có độ phân giải cao hơn từ các tầng high-level semantics

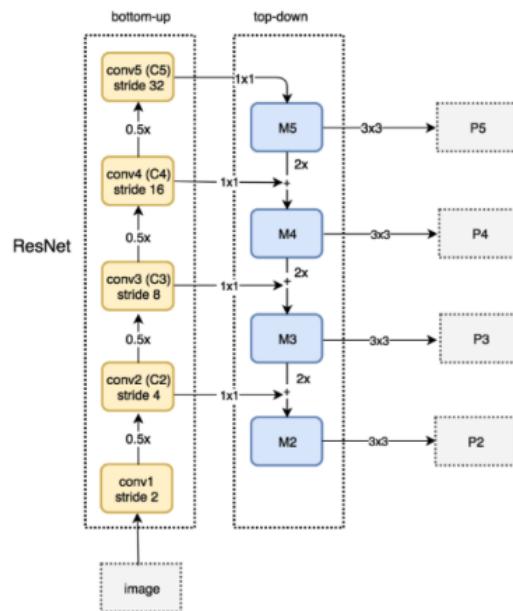
Trong quá trình xây dựng lại các tầng top-down, chúng ta sẽ gặp một vấn đề khá nghiêm trọng là bị mất mát thông tin của các đối tượng. Ví dụ một đối tượng nhỏ khi đi lên (top) sẽ không thấy nó, và từ trên đi ngược lại sẽ không thể tái tạo lại đối tượng nhỏ đó. Để giải quyết vấn đề này, chúng ta sẽ tạo các kết nối (skip connection) giữa các reconstruction layer và các feature maps để giúp quá trình dự đoán các vị trí của đối tượng thực hiện tốt hơn (sao cho hàm mất mát đạt giá trị nhỏ nhất).



Hình 2.28: Top-down

Đồ hình bên dưới biểu diễn chi tiết đường đi theo bottom-up và topdown. P2, P3, P4, P5 là các pyramid của các feature map.

FPN không phải là mô hình phát hiện đối tượng. Nó là mô hình phát hiện đặc



Hình 2.29: Bottom-up và top-down.

trung và được sử dụng trong phát hiện đối tượng. Các feature map từ P2 đến P5 trong Hình 2.29 độc lập với nhau và các đặc trưng được sử dụng để phát hiện đối tượng.

## 2.4.2 Mạng con phân loại

Mạng con phân loại là một mạng tích chập dày đủ (FCN) gắn ở mỗi mức của FPN. Mạng con này chứa 4 lớp tích chập  $3 \times 3$  với 256 bộ lọc, kết hợp với hàm kích hoạt RELU; tiếp đến là lớp tích chập  $3 \times 3$  với số bộ lọc = KA. Do đó đầu ra của feature map có kích thước  $W \times H \times KA$ , trong đó  $W, H$  là các kích thước của feature map, K, A lần lượt là số lượng class và số lượng anchor box. Lý do cho việc sử dụng lớp tích chập cuối cùng của mạng con phân loại có  $K \times A$  bộ lọc vì nếu có “A” anchor box được đề xuất cho mỗi vị trí trong feature map thu được từ lớp tích chập cuối thì mỗi anchor box có khả năng được phân loại được K lớp. Do đó, đầu ra của feature map sẽ có KA bộ lọc.

## 2.4.3 Mạng con hồi quy

Song song với mạng con phân loại đối tượng, mô hình đi kèm một mạng tích chập dày đủ khác vào mỗi cấp của mạng FPN nhằm mục đích điều chỉnh kích thước các anchor box trong việc được phát hiện các đối tượng (nếu có). Cấu trúc của mạng con này giống với mạng con phân loại ở trên, ngoại trừ lớp tích chập cuối có kích thước  $3 \times 3$  với 4 bộ lọc cho kết quả đầu ra feature map của mạng con này có kích thước  $W \times H \times 4A$ . Lý do để lớp tích chập cuối có 4 bộ lọc là vì để xác định vị trí đối tượng trong ảnh, mạng con hồi quy tạo ra 4 giá trị cho mỗi anchor box dự đoán độ lệch tương đối (về tọa độ tâm, chiều rộng và chiều cao) giữa các anchor box với box thực tế chứa đối tượng. Do đó, đầu ra của feature map trong mạng con hồi quy sẽ có 4A bộ lọc (kênh)

## 2.4.4 Hàm mất mát

Trong bài toán nhận diện vật thể, nhãn dữ liệu mô tả chính xác vị trí đối tượng trong bức ảnh được gọi là các ground-truth boxes. Thông qua 2 mạng con ở trên, mỗi anchor box được dự đoán sẽ được phân vào một lớp nào đó và vị trí của anchor box đó trong ảnh. Để tính toán hàm mất mát trong training, ta cần phải so sánh các giá trị dự đoán này với nhãn của dữ liệu. Một anchor box được coi là dự đoán đúng với

một ground truth box nào đó nếu giá trị IOU giữa 2 đối tượng này  $> 0.5$ ; một anchor box được coi là không khớp với bất kỳ ground truth box nào trong ảnh nếu giá trị IOU giữa chúng  $< 0.4$ . Cuối cùng, nếu giá trị IOU giữa một anchor box với bất kỳ ground truth box nào nằm giữa 0.4 và 0.5 sẽ không đóng góp vào hàm mất mát.

Hàm mất mát của mô hình RetinaNet chứa 2 thành phần: Hàm mất mát hồi quy và hàm mất mát phân loại

## Hàm mất mát hồi quy

Giả sử các cặp dữ liệu trùng khớp nhau kí hiệu  $(A^i, G^i)_{i=1,\dots,N}$ , trong đó  $A$  là tập các anchor box,  $G$  là tập các ground truth box,  $N$  là số lượng các ground truth box. Với mỗi anchor, mô hình dự đoán 4 giá trị, kí hiệu lần lượt  $P^i = (P_x^i, P_y^i, P_w^i, P_h^i)$ . Hai giá trị đầu là độ lệch giữa tâm của anchor  $A_i$  và ground truth  $G_i$ , trong đó 2 giá trị cuối là độ lệch giữa chiều cao và chiều rộng của 2 box đó. Tương ứng với mỗi dự đoán, giá trị mục tiêu  $T_i$  được tính như là độ lệch giữa anchor box và gt box theo công thức:

$$T_x^i = (G_x^i - A_x^i)/A_w^i \quad (2.1)$$

$$T_y^i = (G_y^i - A_y^i)/A_h^i \quad (2.2)$$

$$T_w^i = \log(G_w^i/A_w^i) \quad (2.3)$$

$$T_h^i = \log(G_h^i/A_h^i) \quad (2.4)$$

Với các giá trị trên, hàm mất mát hồi quy được định nghĩa là:

$$L_{loc} = \sum_{j \in \{x,y,w,h\}} smooth_{L1}(P_j^i - T_j^i) \quad (2.5)$$

trong đó hàm  $smooth_{L1}(x)$  có công thức:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & |x| \geq 1 \end{cases} \quad (2.6)$$

## Hàm mất mát phân loại

hàm mất mát phân loại mà mô hình retinanet sử dụng là focal loss. Nói một cách đơn giản, Focal loss (FL) là một phiên bản cải tiến của hàm Cross-Entropy Loss (CE)

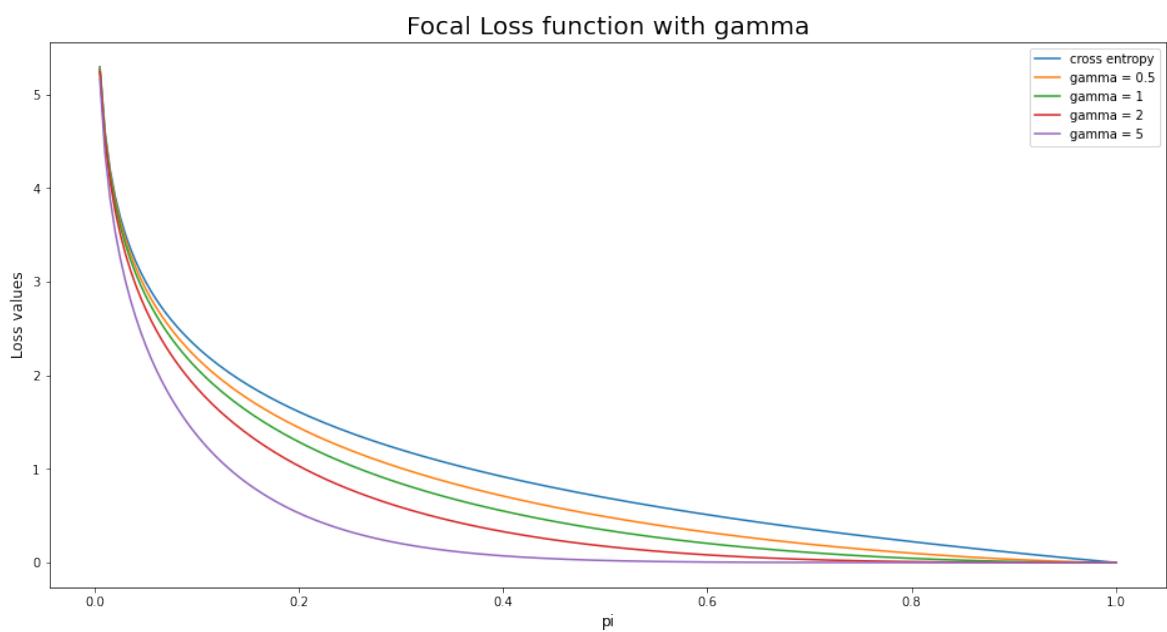
cố gắng xử lý vấn đề mất cân bằng giữa các lớp foreground và background; một điều thường xảy ra trong các bài toán object detection .Hàm được xây dựng bằng cách gán nhiều trọng số hơn cho các mẫu khó hoặc dễ bị phân loại sai và giảm trọng số cho các mẫu dễ phân loại. Vì vậy, Focal loss làm giảm sự đóng góp tổn thất từ các mẫu dễ phân loại và tăng tầm quan trọng của việc sửa lỗi các mẫu được phân loại sai Đầu tiên ta sẽ tìm hiểu về hàm Cross-Entropy (CE) loss trong bài toán phân loại nhị phân:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1 - p) & \text{otherwise,} \end{cases} \quad (2.7)$$

với  $y$  là nhãn phân loại của dữ liệu,  $p \in [0, 1]$  là xác suất dự đoán cho lớp có nhãn  $y$  của mô hình. Để thuận tiện, ta định nghĩa  $p_t$ :

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases} \quad (2.8)$$

Khi đó  $CE(p, y) = CE(p_t) = -\log(p_t)$



Hình 2.30: Đồ thị hàm focal loss với các giá trị  $\gamma$  khác nhau

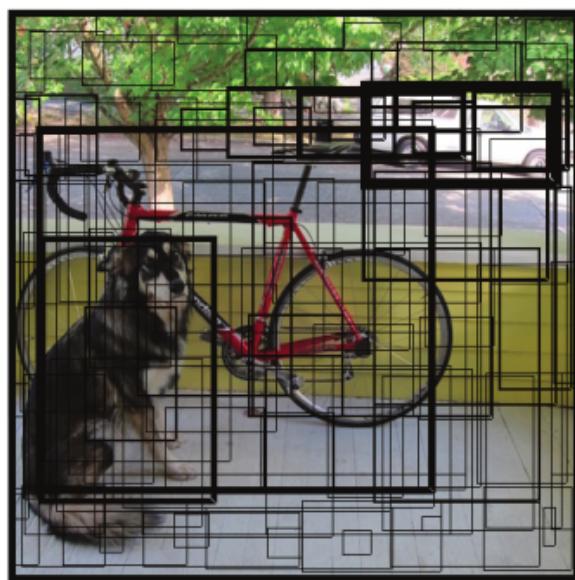
Dồ thị của CE loss có thể được xem là đường cong màu xanh trong hình 2.31 Một điểm đáng chú ý có thể dễ dàng nhận ra từ đồ thị trên, đó là ngay cả những mẫu dễ phân loại (có  $p_t >> 5$ ) cũng có giá trị loss khá lớn.

Trong cross entropy ta thấy rằng vai trò đóng góp vào loss function của các class cùng bằng  $-\log(p_i)$ . Khi xảy ra hiện tượng mất cân bằng, chúng ta muốn rằng mô hình sẽ dự báo chuẩn hơn đối với những class thiểu số. Do đó cần một hàm loss function hiệu quả hơn, có thể điều chỉnh được giá trị phạt lớn hơn đối với nhóm thiểu số.

Một kỹ thuật thường dùng để giải quyết việc mất cân bằng giữa các lớp là áp dụng trọng số  $\alpha$  bằng nghịch đảo tần suất nhãn vào hàm cross entropy. Hàm loss function mới được gọi là balanced cross entropy:

$$BCE(\mathbf{p}, \mathbf{q}) = -\alpha_i \log(q_i), \quad \text{với } p_i = 1 \quad (2.9)$$

hàm mất mát này là một mở rộng nhỏ của CE loss mà tác giả của bài báo coi là cơ sở cho đề xuất hàm mất mát mới: Focal loss. Focal loss là hàm loss function lần đầu được giới thiệu trong RetinaNet. Hàm loss function này đã chứng minh được tính hiệu quả trong các bài toán object detection. Đây là lớp bài toán có sự mất cân bằng nghiêm trọng giữa hai class positive (các bounding box có chứa object) và negative (các bounding box không chứa object). Thường thì negative có số lượng lớn hơn positive rất nhiều. Lấy ví dụ như hình bên dưới :



Hình 2.31: Minh họa các lớp positive và negative trong object detection

# Chương 3

## Ứng dụng học sâu vào bài toán đếm cây trên ảnh viễn thám

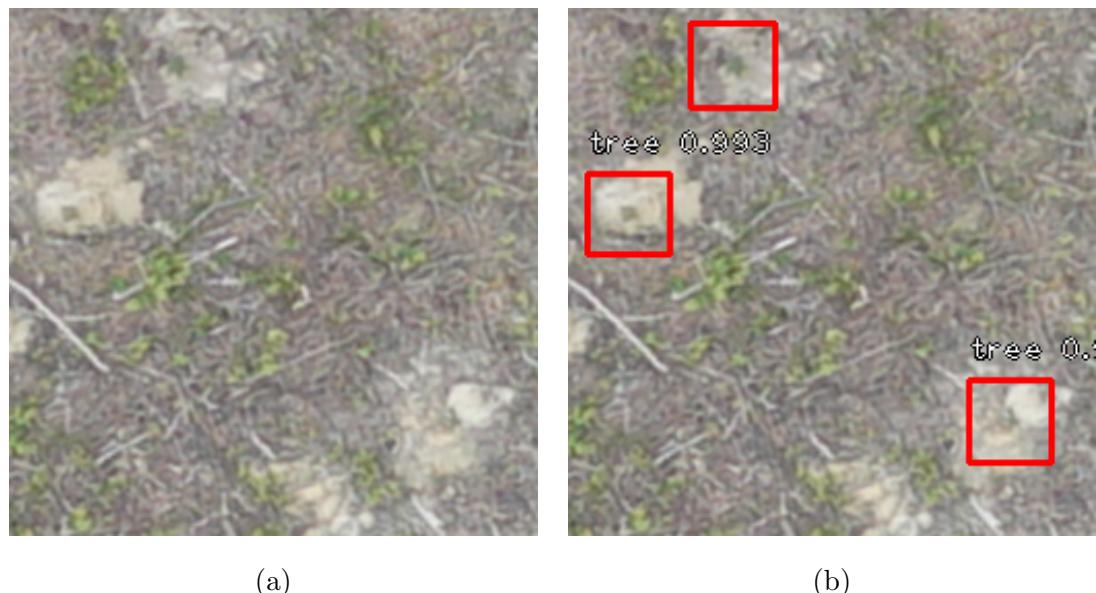
### 3.1 Giới thiệu bài toán

Cây keo là một loại cây dễ trồng, có khả năng thích nghi với các loại đất nghèo chất dinh dưỡng, ở những nơi thời tiết khắc nghiệt. Gỗ keo có rất nhiều tác dụng nên keo rất phù hợp với các dự án lâm nghiệp thương mại: là nguyên liệu sử dụng để sản xuất đồ nội thất xuất khẩu và sản xuất giấy. Cây keo ở Việt Nam ngày càng được trồng với quy mô lớn, nhằm nhanh chóng phủ xanh đồi núi trọc keo được trồng thành rừng, đem lại nguồn thu nhập lớn, tạo công ăn việc làm cho nhiều người. Để đáp ứng nhu cầu tiêu thụ gỗ ngày càng tăng ở trong nước và thế giới, sản lượng nông nghiệp tiêu dùng, năng suất cây trồng phải được ước tính. Việc phát hiện, kiểm soát cây keo trên một diện tích lớn trong khu rừng giúp cho doanh nghiệp và các nhà quản lý biết được tình trạng hiện tại của các cây keo, có thể trồng thêm được cây nào vào những khu đất còn trống. Tuy nhiên việc thu thập thủ công dữ liệu cây trồng để lưu trữ hồ sơ trên một diện tích đất lớn là không khả thi đối với con người, gây tốn kém tiền bạc và dễ bị ảnh hưởng bởi lỗi của con người.

Một tiềm năng lớn để giải quyết vấn đề này là sử dụng các hình ảnh vệ tinh kết hợp với các thuật toán thị giác máy tính để phát hiện quản lý những cây trồng này. Sử dụng hình ảnh vệ tinh để khảo sát sẽ giảm bớt một phần công việc của con người

khi phải trực tiếp đến khu vực đó tiến hành theo dõi, phân tích. Các bức ảnh viễn thám được chụp lại với kích thước lớn, có thể bao quát một khu diện tích đất trống rộng giúp tiết kiệm thời gian điều tra khảo sát, từ đó giúp các nhà quản lý của doanh nghiệp có thể nắm bắt tình hình kịp thời, lập kế hoạch sản xuất phù hợp.

Để việc giải quyết bài toán này trở nên đơn giản và tự động hơn, trong Đồ án này, em trình bày về việc sử dụng các phương pháp Deep Learning để tiến hành phân tích.



## 3.2 Mô hình hóa bài toán và thiết kế dữ liệu luyện

a) Mô hình hóa bài toán

**Input:**

Ảnh viễn thám được lưu ở định dạng .tif, có kích thước lớn (từ vài chục MB đến GB)

**Tiền xử lý:**

Ảnh đầu vào sẽ được chia nhỏ thành các ảnh có kích thước  $256 \times 256$  (bằng với kích thước trong tập ảnh training), với tỉ lệ *overlap* giữa các tấm ảnh là 0.5

## Output:

Các file định dạng txt chứa các bounding box mà mô hình đoán đó là vật thể trong bức ảnh đó. Vì mục tiêu của bài toán là nhận dạng và đếm số cây trên một bức ảnh lớn, nên ta sẽ ghép các tấm ảnh nhỏ ở trên lại để đưa về hình ảnh gốc của chúng. Vì trong quá trình tiền xử lý, ta thiết lập overlap = 0.5 nên việc các bounding box của các bức ảnh gần nhau sẽ đè lên nhau, gây ra hiện tượng trùng lặp. Để giải quyết vấn đề này, các bounding box thu được trên toàn bộ các ảnh đầu tiên sẽ được đưa về tọa độ địa lý, sau đó sử dụng thuật toán NMS để loại bỏ các bounding box trùng nhau (nếu như giá trị IOU của hai bounding box này vượt quá một ngưỡng (trong mô hình này em lấy ngưỡng = 0.3)). Kết quả sau khi thực hiện thuật toán NMS sẽ

### b) Thiết kế dữ liệu luyện mạng

Dữ liệu sử dụng trong bài toán này được cung cấp bởi thầy giáo hướng dẫn TS. Lê Hải Hà. Dữ liệu bao gồm 2 thư mục, mỗi thư mục gồm một vài ảnh ở định dạng GeoTif và các **shapefile** chứa tọa độ các bounding box trong các bức ảnh đó

W03_202003311249_RI_RSK_RSKA014702_RGB		21:13, 12-11-2 Reiwa	95,7 MB	Thư mục
W03_202003311249_RI....RSKA014702_RGB_0.tif		15:39, 04-11-2 Reiwa	23,7 MB	Ảnh TIFF
W03_202003311249_RI....K_RSKA014702_RGB_1.tif		15:39, 04-11-2 Reiwa	23,7 MB	Ảnh TIFF
W03_202003311249_RI....RSKA014702_RGB_2.tif		15:39, 04-11-2 Reiwa	23,7 MB	Ảnh TIFF
W03_202003311249_RI....RSKA014702_RGB_3.tif		15:39, 04-11-2 Reiwa	23,7 MB	Ảnh TIFF
W03_202003311249_RI_RSK_RSKA014702_RGB.dbf		09:24, 11-08-2 Reiwa	9 KB	Tài liệu
W03_202003311249_RI_RSK_RSKA014702_RGB.prj		09:24, 11-08-2 Reiwa	389 byte	Tài liệu
W03_202003311249_RI_RSK_RSKA014702_RGB.shp		09:24, 11-08-2 Reiwa	857 KB	ESRI Sh...cument
W03_202003311249_RI_RSK_RSKA014702_RGB.shx		09:24, 11-08-2 Reiwa	6 KB	Tài liệu
W05_202003281250_RI_RSK_RSKA003603_RGB		00:32, Hôm nay	292,2 MB	Thư mục
W05_202003281250_RI....RSKA003603_RGB_0.tif		15:39, 04-11-2 Reiwa	48,1 MB	Ảnh TIFF
W05_202003281250_RI....RSKA003603_RGB_1.tif		15:38, 04-11-2 Reiwa	48,1 MB	Ảnh TIFF
W05_202003281250_RI....RSKA003603_RGB_2.tif		15:39, 04-11-2 Reiwa	48,1 MB	Ảnh TIFF
W05_202003281250_RI....RSKA003603_RGB_3.tif		15:39, 04-11-2 Reiwa	48,1 MB	Ảnh TIFF
W05_202003281250_RI....RSKA003603_RGB_4.tif		15:39, 04-11-2 Reiwa	48,1 MB	Ảnh TIFF
W05_202003281250_RI....RSKA003603_RGB_5.tif		15:39, 04-11-2 Reiwa	48,1 MB	Ảnh TIFF
W05_202003281250_RI_RSK_RSKA003603_RGB.dbf		09:25, 11-08-2 Reiwa	42 KB	Tài liệu
W05_202003281250_RI_RSK_RSKA003603_RGB.prj		09:25, 11-08-2 Reiwa	389 byte	Tài liệu
W05_202003281250_RI....RSKA003603_RGB.shp		09:25, 11-08-2 Reiwa	3,8 MB	ESRI Sh...cument
W05_202003281250_RI_RSK_RSKA003603_RGB.shx		09:25, 11-08-2 Reiwa	28 KB	Tài liệu

Hình 3.2: Dữ liệu sử dụng trong bài toán



Hình 3.3: Minh họa ảnh viễn thám với phần mềm QGIS



Hình 3.4: Toạ độ các bounding box được gán nhãn trên ảnh, biểu thị bởi các hình tròn màu vàng

Các ảnh trong bộ dữ liệu ở trên có 2 loại kích thước :  $3141 \times 1885$  và  $3874 \times 3100$ .

Để thuận tiện cho việc training mô hình, ta sẽ chia nhỏ mỗi ảnh thành nhiều ảnh con có kích thước  $256 \times 256$  với giá trị  $overlap = 0.5$ . Để lấy bounding box trong từng bức ảnh được cắt ra, đầu tiên ta tiến hành đọc file *shapefile* của ảnh đó bằng thư viện *geopandas* (hình 3.5):

	<b>FID</b>	<b>geometry</b>
0	0	POLYGON ((782749.090 87317.692, 782749.088 873...
1	1	POLYGON ((782748.127 87316.369, 782748.126 873...
2	2	POLYGON ((782736.356 87317.697, 782736.355 873...
3	3	POLYGON ((782741.586 87317.399, 782741.584 873...
4	4	POLYGON ((782747.997 87314.795, 782747.995 873...
...	...	...
3462	3462	POLYGON ((782858.704 87068.587, 782858.702 870...
3463	3463	POLYGON ((782858.220 87067.431, 782858.218 870...
3464	3464	POLYGON ((782909.154 87222.547, 782909.152 872...
3465	3465	POLYGON ((782913.085 87225.018, 782913.083 872...
3466	3466	POLYGON ((782958.935 87267.731, 782958.933 872...
3467	rows × 2 columns	

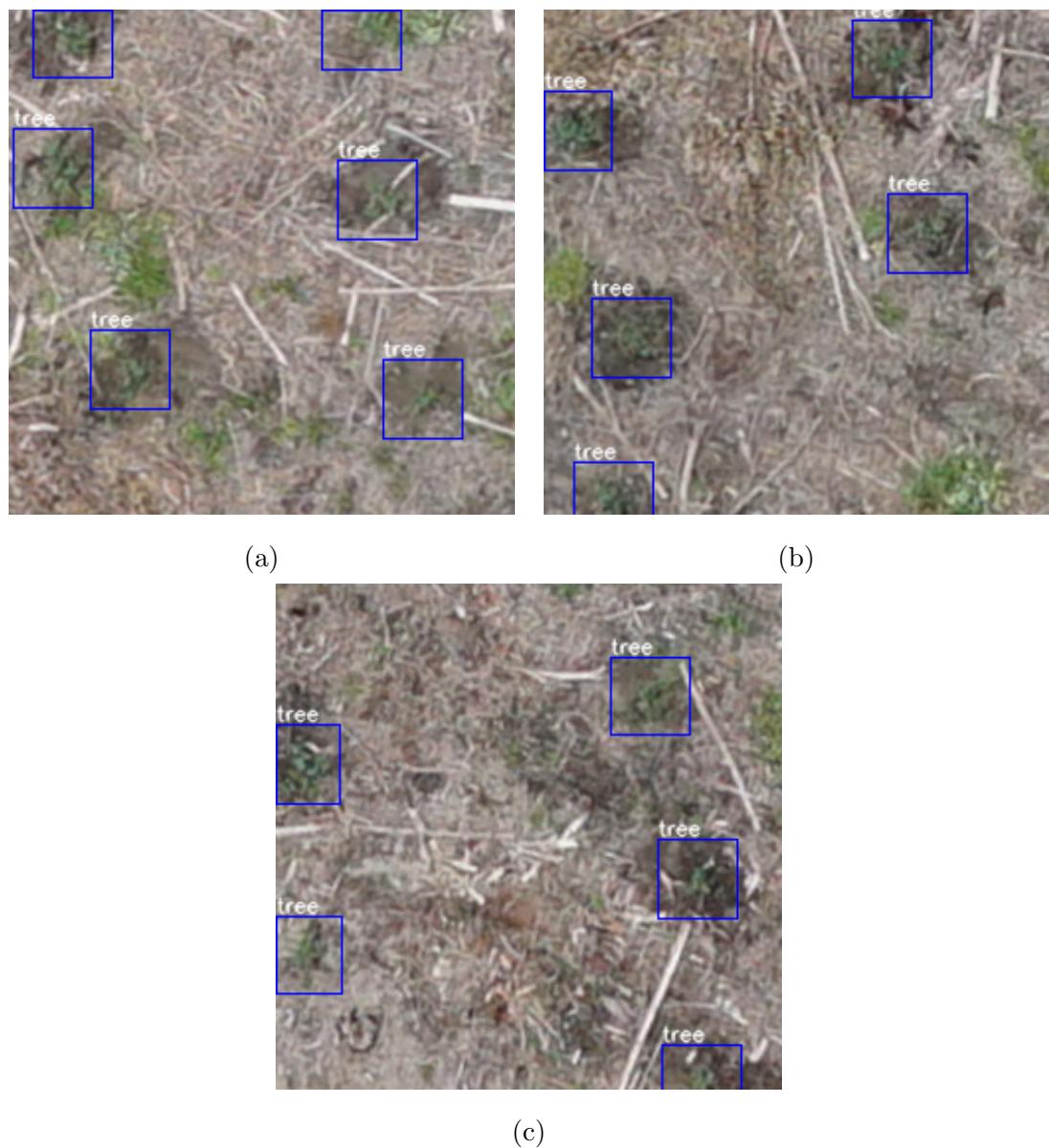
Hình 3.5: Đọc file shapefile với thư viện geopandas

Vì tọa độ các câu được đánh dấu theo hình tròn nên ta cần chuyển định dạng hình tròn về bounding box (hình 3.6)

polygons.bounds				
	minx	miny	maxx	maxy
0	782748.289193	87317.294211	782749.089590	87318.089250
1	782747.327019	87315.971222	782748.127415	87316.766261
2	782735.556070	87317.299441	782736.356466	87318.094480
3	782740.785279	87317.001376	782741.585675	87317.796415
4	782747.196289	87314.397230	782747.996685	87315.192269
...	...	...	...	...
3462	782857.903169	87068.189629	782858.703566	87068.984669
3463	782857.419266	87067.032845	782858.219664	87067.827885
3464	782908.353211	87222.149652	782909.153608	87222.944692
3465	782912.284263	87224.619968	782913.084660	87225.415008
3466	782958.134333	87267.332957	782958.934730	87268.127997
3467 rows × 4 columns				

Hình 3.6: Toạ độ 4 đỉnh của một polygon

Tuy nhiên trong hình trên, mỗi điểm là một toạ độ trên một mặt phẳng địa lý, ta phải chuyển các toạ độ  $x, y$  về toạ độ cột, hàng tương ứng. Một vài kết quả thu được sau khi tiến hành cắt ảnh và lấy bounding box tương ứng cho mỗi bức ảnh đó:



Hình 3.7: Main caption for subfigures a, b and c

Kết quả sau khi cắt ảnh ra ta thu được bộ dữ liệu hơn 5000 ảnh kích thước  $256 \times 256$  cùng với tọa độ các bounding box tương ứng trong từng ảnh đó

### 3.3 Huấn luyện mạng neural

Chúng ta sẽ tiến hành huấn luyện mô hình dựa trên bộ dữ liệu được tạo ra ở trên. Mô hình mà em sử dụng trong bài toán này là Faster- RCNN; trong đó backbone sử dụng

ở đây là ResNet101 phục vụ cho việc trích xuất các đặc trưng trong ảnh

### **3.4 Xây dựng chương trình**

# **Chương 4**

## **Cài đặt chương trình và đánh giá kết quả**

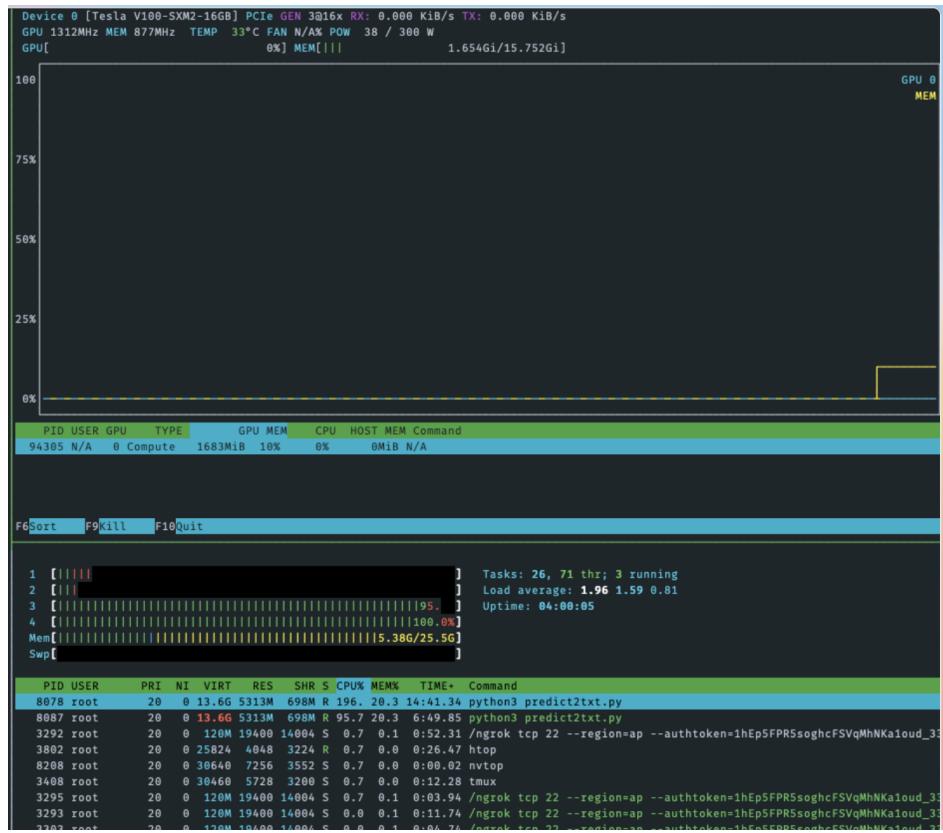
### **4.1 Môi trường cài đặt chương trình và các yêu cầu liên quan**

#### **4.1.1 Môi trường cài đặt chương trình**

Chương trình được cài đặt trên ngôn ngữ Python (phiên bản 3.6.9) và được thử nghiệm trên hệ điều hành máy ảo của google colab sử dụng chip 4 cores , bộ nhớ RAM 24GB, card đồ họa NVIDIA Tesla V100 16GB (hình 4.1)

#### **4.1.2 Các yêu cầu liên quan**

- Ngôn ngữ sử dụng: Python phiên bản 3.6.9
- IDE sử dụng: Visual Studio Code
- Các thư viện Python sử dụng: pandas, numpy, matplotlib, torch, os , cv2 và xlwt.



Hình 4.1: Cấu hình máy tính sử dụng cho bài toán

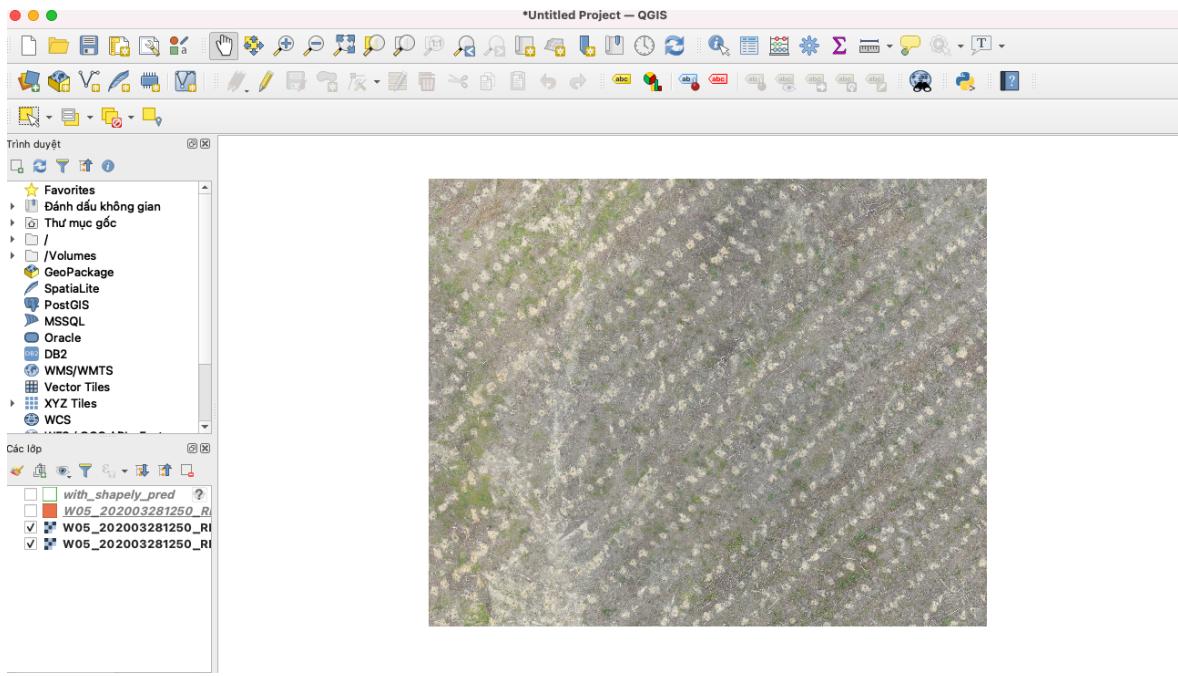
## 4.2 Dữ liệu đầu vào

Vì dữ liệu em được cung cấp bị hạn chế, nên trong bộ dữ liệu ban đầu em lấy ra 1 ảnh chưa được xử lý cho tập train và validation (hình 4.2) để làm dữ liệu đầu vào cho mô hình nhận diện vật thể đã xây dựng ở chương trước. Lý do của việc lấy một ảnh kích thước lớn để dự đoán là vì theo yêu cầu của khách hàng, họ cần đánh giá kết quả mô hình trên toàn bộ diện tích đất của họ chứ không phải các tấm ảnh kích thước  $256 \times 256$  được cắt nhỏ từ tấm ảnh ban đầu.

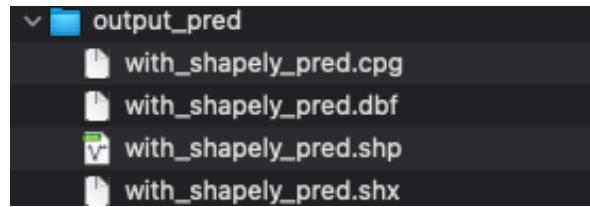
## 4.3 Kết quả mô hình

Thời gian chạy chương trình cho ảnh trên: 43 giây

Hình ảnh kết quả được lưu lại ở dạng shapefile: (hình 4.3)



Hình 4.2: Dữ liệu đầu vào để phát hiện vật thể

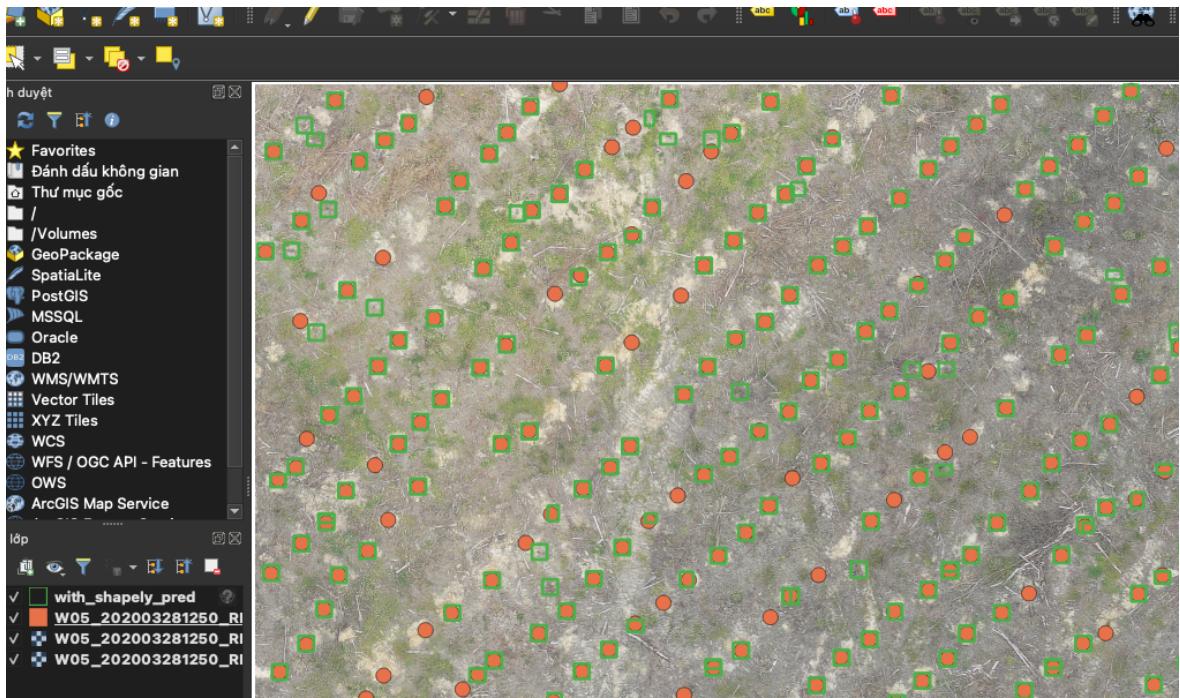


Hình 4.3: Kết quả mô hình: thư mục chứa các shapfile

Ta sử dụng file shapefile ở trên để mô phỏng kết quả trên phần mềm QGIS (hình 4.4):

## 4.4 Đánh giá kết quả

Hình 4.4 ở trên chỉ cho ta một cách khái quát về kết quả mô hình khi so sánh với các bounding box ban đầu. Để đánh giá kết quả của mô hình, trước tiên ta định nghĩa 2 bounding box cùng phát hiện một đối tượng nếu giá trị IOU của chúng  $\geq 0.3$ . Khi đó, bounding box được dự đoán từ mô hình sẽ được coi là một *True positive*, ngược lại sẽ là *False positive*. Trong trường hợp có nhiều hơn 1 bounding box cùng dự đoán một đối tượng, ta sẽ lấy bounding box có confidence (độ tin cậy) cao nhất làm TP, các



Hình 4.4: Kết quả mô hình: các ô vuông màu xanh là kết quả phát hiện của mô hình; chấm màu cam là các cây được gán nhãn sẵn.

bounding box còn lại sẽ là FP. Để so sánh một cách khách quan, em sử dụng cùng một tập các thang đo phổ biến là Precision, Recall và F1-score. Trong đó, Precision là tỉ lệ giữa số tên miền phân loại chính xác trên tổng số tên miền được dự đoán của mỗi lớp. Recall là tỉ lệ giữa số tên miền được phân loại đúng theo một nhãn trên tổng số tên miền được gán theo nhãn đó. F1-score là trung bình điều hòa giữa hai giá trị Precision và Recall. Các tham số trên được biểu diễn trong các biểu thức:

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

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

$$F1 - score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

```
Precision: tree: 0.8658536585365854
Recall: tree: 0.8214876033057851
F1-score: tree: 0.8430873621713316
-----Done-----
```

Hình 4.5: Đánh giá kết quả

## 4.5 Định hướng phát triển trong tương lai

Từ kết quả thực nghiệm cho thấy, chương trình đã có những thành công nhất định. Song bên cạnh đó cũng còn khá nhiều nhược điểm cần cải tiến. Trong quá trình hoàn thành đồ án em đã đầu tư nhiều thời gian và công sức với bài toán này, và nhận thấy đây là bài toán có khả năng phát triển cao hơn nữa . Em xin đưa ra một số hướng phát triển tiếp theo cho bài toán như sau:

- Phát triển và cài đặt giao diện chương trình để có thể dễ dàng chạy và sử dụng hơn.
- Sử dụng thêm một vài phép xử lý ảnh để tăng cường dữ liệu phục vụ cho việc training
- Cải tiến, thay đổi các thuật toán để loại bỏ các bounding box trùng nhau nhanh hơn khi sử dụng thuật toán NMS; giúp chương trình chạy tối ưu hơn

# **Tài liệu tham khảo**

## **Tài liệu tham khảo tiếng Việt**

- [1.] Nguyễn Đức Nghĩa, Nguyễn Tô Thành, "Toán rời rạc", Nhà xuất bản Đại Học Quốc Gia Hà Nội, 1997, tr. 147-155.
- [2.] Tống Đinh Quỳ, "Giáo trình xác suất thống kê", Nhà xuất bản Bách Khoa-Hà Nội, 2016.

## **Tài liệu tham khảo tiếng Anh**

- [1.] Martin Aigner, Günter M. Ziegler, “Proofs from THE BOOK”, 6<sup>th</sup> Edition.
- [2.] Valle Martinez, Vicente, "Notes on the proof of the van der Waerden permanent conjecture"(2018).