

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
CHƯƠNG TRÌNH KỸ SƯ CHẤT LƯỢNG CAO VIỆT PHÁP
KHOA ĐIỆN - ĐIỆN TỬ
BỘ MÔN VIỄN THÔNG



LUẬN VĂN TỐT NGHIỆP

**HỆ THỐNG GIÁM SÁT VIỆC SỬ
DỤNG THIẾT BỊ BẢO HỘ CÁ
NHÂN ỨNG DỤNG MẠNG HỌC
SÂU**

NGUYỄN THÁI SƠN - 1512847

**GIẢNG VIÊN HƯỚNG DẪN:
PGS. TS. HÀ HOÀNG KHA**

Thành phố Hồ Chí Minh, tháng 6 năm 2020

LỜI CẢM ƠN

Khoảng thời gian học tập và rèn luyện tại **Trường Đại học Bách Khoa Thành phố Hồ Chí Minh** đã trang bị cho em rất nhiều kiến thức hữu ích và cần thiết về chuyên môn và xã hội để em có thể trở thành một công dân tốt và một kỹ sư có năng lực. Con đường học tập ở đại học trong suốt 5 năm vừa qua là không hề dễ dàng với muôn vàn thử thách và khó khăn. Để vượt qua những rào cản ấy, bên cạnh sự cố gắng của bản thân em còn là sự ủng hộ và giúp đỡ tận tình của quý **Thầy Cô, Gia đình** và **Bạn bè**. Em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến quý **Thầy Cô Trường Đại học Bách Khoa Thành phố Hồ Chí Minh**, quý **Thầy Cô Khoa Điện – Điện Tử**, những người đã đi cùng với tri thức và tâm huyết truyền đạt vốn kiến thức quý báu của mình cho chúng em.

Em muốn giành riêng lời cảm ơn đặc biệt cho Thầy hướng dẫn của mình – PGS. TS. Hà Hoàng Kha – người đã luôn tận tình hướng dẫn và giúp đỡ em trong quá trình thực hiện luận văn này.

Ngoài ra, em cũng muốn giành cho gia đình và bạn bè lời cảm ơn chân thành và đặc biệt là ba mẹ và ông bà em, những người đã luôn là chỗ dựa tinh thần vững chắc cho em trong những thời điểm khó khăn.

Trong quá trình thực hiện luận văn chắc chắn không thể tránh khỏi những sai sót, em rất mong nhận được những ý kiến đóng góp quý báu của quý Thầy Cô để em có thể học hỏi thêm những điều tốt đẹp và hoàn thiện luận văn của mình. Sau cùng, em xin kính chúc quý **Thầy Cô Trường Đại học Bách Khoa Thành phố Hồ Chí Minh** và **Thầy Hà Hoàng Kha** dồi dào sức khỏe, đạt được nhiều thành công trong cuộc sống và luôn giữ vững niềm đam mê nghiên cứu và giảng dạy để có thể tiếp tục truyền lửa tri thức cho những thế hệ sau.

Thành phố Hồ Chí Minh, tháng 6 năm 2020

Nguyễn Thái Sơn

TÓM TẮT LUẬN VĂN

Ý tưởng về việc kết hợp trí tuệ nhân tạo và thị giác máy tính để ứng dụng vào các bài toán giám sát đã xuất hiện từ lâu. Tuy nhiên chỉ đến những năm gần đây khi các thiết bị phần cứng có thể đáp ứng được yêu cầu về tính kỹ thuật và kinh tế thì các ứng dụng sử dụng các công nghệ này mới dần trở nên phổ biến. Một trong những ứng dụng đang rất được quan tâm là đảm bảo an toàn lao động của công nhân xây dựng thông qua một hệ thống giám sát sử dụng camera và trí tuệ nhân tạo để theo dõi việc sử dụng các thiết bị bảo hộ lao động của những người làm việc trong công trường.

Trong khuôn khổ của luận văn này, hệ thống giám sát việc sử dụng thiết bị bảo hộ cá nhân sẽ tập trung vào ba thiết bị thường gặp trong công trường xây dựng: mũ cứng, áo dạ quang bảo hộ và khẩu trang. Hệ thống sử dụng phương pháp phát hiện và phân loại vật thể YOLO, được xây dựng trên cơ sở mạng tích chập - CNN. Khi hoạt động, hệ thống sẽ có thể phát hiện và phân loại việc sử dụng các thiết bị bảo hộ lao động của công nhân ở công trường. Nếu phát hiện một trường hợp không sử dụng thiết bị bảo hộ lao động đang được theo dõi thì hệ thống sẽ gửi cảnh báo cho quan sát viên để kiểm tra và nhắc nhở. Việc này sẽ hỗ trợ rất nhiều trong công tác đảm bảo an toàn lao động trong xây dựng.

Hệ thống được viết bằng ngôn ngữ Python, mô hình máy học YOLOv3 và thư viện thị giác máy tính OpenCV.

Mục lục

| | | |
|----------|--|-----------|
| 1 | Giới thiệu | 9 |
| 1.1 | Đặt vấn đề | 9 |
| 1.2 | Mục tiêu nghiên cứu | 10 |
| 1.3 | Phạm vi nghiên cứu | 10 |
| 1.4 | Phương pháp nghiên cứu | 10 |
| 1.5 | Cấu trúc luận văn | 11 |
| 2 | Cơ sở lý thuyết | 12 |
| 2.1 | Gradient Descent | 12 |
| 2.1.1 | Batch Gradient Descent | 14 |
| 2.1.2 | Stochastic Gradient Descent | 15 |
| 2.1.3 | Mini-batch Gradient Descent | 16 |
| 2.1.4 | Điều kiện dừng của giải thuật | 17 |
| 2.2 | Backpropagation | 17 |
| 2.3 | Mạng neuron tích chập | 21 |
| 2.3.1 | Lớp tích chập | 21 |
| 2.3.2 | Lớp pooling | 26 |
| 2.3.3 | Lớp đầy đủ kết nối | 26 |
| 2.3.4 | Mô hình mạng neuron tích chập | 27 |
| 2.4 | YOLOv3 | 27 |
| 2.4.1 | Unified Detection | 28 |
| 2.4.2 | Kiến trúc mạng YOLOv3 | 31 |
| 3 | Phương pháp tiếp cận | 34 |
| 3.1 | Xây dựng tập dữ liệu | 34 |
| 3.1.1 | Xác định yêu cầu bài toán | 34 |
| 3.1.2 | Thu thập hình ảnh và dán nhãn | 35 |
| 3.2 | Huấn luyện mạng YOLOv3 sử dụng framework Darknet | 36 |

| | | |
|---|-------------------|----|
| 4 | Phân tích kết quả | 43 |
| 5 | Kết luận | 44 |

Danh sách hình vẽ

| | | |
|-----|---|----|
| 2.1 | Cực tiểu địa phương (màu xanh) và cực tiểu toàn cục (màu đỏ) | 13 |
| 2.2 | Batch Gradient Descent với bài toán hồi quy tuyến tính. Toàn bộ số điểm đầu vào đều được dùng để cập nhật các vector trọng số (a, b) cho đường hồi quy tại mỗi bước, với a là độ dốc và b độ sai lệch. | 14 |
| 2.3 | Stochastic Gradient Descent với bài toán hồi quy tuyến tính. Một điểm đầu vào được chọn ngẫu nhiên để cập nhật các vector trọng số (a, b) cho đường hồi quy tại mỗi iteration, với a là độ dốc và b độ sai lệch. | 15 |
| 2.4 | Mini-batch Gradient Descent với bài toán hồi quy tuyến tính. Một batch sẽ gồm ba điểm đầu vào được chọn ngẫu nhiên để cập nhật các vector trọng số (a, b) cho đường hồi quy tại mỗi iteration, với a là độ dốc và b độ sai lệch. Một epoch sẽ gồm mười batch. | 16 |
| 2.5 | Mô hình mạng neuron đơn giản. | 18 |
| 2.6 | Mô hình mạng neuron tích chập đơn giản. Lớp nhận hình ảnh vào màu đỏ là một lớp có cấu trúc ba chiều với chiều rộng và chiều cao là chiều rộng và chiều cao của hình ảnh đầu vào, chiều sâu bằng ba ứng với ba kênh màu đỏ, xanh lá và xanh dương. Các lớp của mạng neuron tích chập sẽ chuyển đổi một nhóm các ma trận thành một nhóm các ma trận khác. Lớp ngoài cùng là lớp phân loại, có kích thước các chiều tương ứng với một vector. | 22 |
| 2.7 | Hình ảnh đầu vào gồm ba kênh màu được mô hình hóa thành tensor với chiều cao và chiều rộng là chiều cao và chiều rộng của ảnh, chiều sâu là ba. | 22 |

| | | |
|------|---|----|
| 2.8 | Hình ảnh sau khi được đưa qua đầu vào và chuyển đổi thành dữ liệu ba chiều sẽ được đưa vào lớp convolution đầu tiên. Một kernel có kích thước $3 \times 3 \times 3$ (góc trên bên trái của mô hình ngoài cùng bên phải) được trượt qua hình đầu vào. | 23 |
| 2.9 | Ví dụ về phép toán của sổ trượt với kích thước 3×3 | 23 |
| 2.10 | Bên trái, ma trận 3×3 được zero padding với <i>padding</i> = 1. Bên phải, ma trận 3×3 được zero padding với <i>padding</i> = 2 . | 24 |
| 2.11 | Ví dụ về một kernel có kích thước $3 \times 3 \times 3$ | 24 |
| 2.12 | Ví dụ về phép toán của một kernel lên một vị trí của ảnh trong lớp tích chập. | 25 |
| 2.13 | Một lớp tích chập có <i>k</i> kernel với kích thước $3 \times 3 \times 3$, <i>stride</i> = 1, <i>padding</i> = 1. Đầu vào là một tensor có kích thước $h \times w \times d$ đầu ra của phép tích chập lên tensor này khi khối tích chập có các thông số ở trên là một tensor có kích thước $h \times w \times k$. | 26 |
| 2.14 | Bên trái, lớp pooling với cửa sổ trượt lấy giá trị lớn nhất với kích thước cửa sổ 2×2 , <i>stride</i> = 1, <i>padding</i> = 0. Bên phải, lớp pooling với cửa sổ trượt lấy giá trị trung bình với kích thước cửa sổ 2×2 , <i>stride</i> = 2, <i>padding</i> = 0. | 27 |
| 2.15 | Mạng neuron tích chập gồm hai lớp tích chập và pooling, một lớp kết nối đầy đủ. | 27 |
| 2.16 | Hình ảnh được chia thành mạng lưới ô vuông $S \times S$ | 28 |
| 2.17 | Miêu tả việc tính toán IOU. | 29 |
| 2.18 | Mô hình dự đoán bounding box của YOLO. | 30 |
| 2.19 | Kiến trúc mạng Darknet-53. | 32 |
| 3.1 | (1) Mũ bảo hộ, (2) Áo bảo hộ, (3) Khẩu trang | 34 |
| 3.2 | Kết quả nhận dạng mong muốn. | 35 |
| 3.3 | Định dạng nhãn của YOLO. | 36 |
| 3.4 | Thống kê số lượng vật thể ứng với từng class. Wearing a hardhat: 27565, Not wearing a hardhat: 45888, Wearing a safety vest: 10264, Not wearing a safety vest: 52996, Wearing a mask: 12557, Not wearing a mask: 46501. | 37 |

Danh sách bảng

| | | |
|-----|--|----|
| 2.1 | So sánh hiệu năng của Darknet-53 với các mạng khác. Accuracy, Bn Ops - billions of operations, BFLOP/s - billion floating point operations per second, và FPS - frames per second. | 33 |
|-----|--|----|

Danh sách từ viết tắt

MRC Maximal Ratio Combining

Chương 1

Giới thiệu

1.1 Đặt vấn đề

Ngành xây dựng luôn được coi là một trong những ngành ẩn chứa nhiều rủi ro về tai nạn lao động và khả năng mắc các bệnh nghề nghiệp. Trên thực tế nhiều vụ tai nạn nghiêm trọng đã xảy ra, lấy đi sinh mạng hoặc để lại những thương tật nặng nề cho người lao động khiến họ mất khả năng làm việc, sinh hoạt như người bình thường. Kéo theo đó là nỗi đau về tinh thần và gánh nặng về kinh tế cho những thành viên trong gia đình người bị nạn. Do đó vấn đề đảm bảo an toàn vệ sinh lao động luôn là một trong những vấn đề được quan tâm hàng đầu trong ngành xây dựng.

Một trong những nguyên nhân chính gây ra những vụ tai nạn thương tâm là việc người lao động không sử dụng trang thiết bị bảo hộ cá nhân trong quá trình lao động. Vấn đề này không chỉ xuất phát từ sự chủ quan của cá nhân người lao động mà còn ở sự thiếu sót, lỏng lẻo trong quá trình giám sát công trình của nhà thầu và người sử dụng lao động. Đối với người lao động, những điều kiện khắc nghiệt của môi trường làm việc như nhiệt độ ngoài trời cao hay thường xuyên phải vận động mạnh khiến đổ mồ hôi liên tục đã khiến họ chấp nhận đánh đổi sự an toàn của bản thân để đổi lấy sự thoải mái. Còn đối với những người giám sát công trình, họ không thể bao quát được toàn bộ quá trình làm việc tại các nơi làm việc khác nhau, do đó không thể nhắc nhở người lao động kịp thời trước khi xảy ra những tai nạn mà hậu quả là có thể tránh khỏi hoặc được giảm nhẹ nếu người lao động có sử dụng trang thiết bị bảo hộ cá nhân.

Để tăng cường năng lực thực hiện đảm bảo an toàn vệ sinh lao động ở các

công trình, nhiều chủ đầu tư và nhà thầu đã tiến hành lắp đặt các hệ thống camera giám sát quá trình làm việc. Các hệ thống này giúp giám sát viên có thể quan sát nhiều vị trí một lúc mà không cần phải di chuyển qua các địa điểm khác nhau trong công trình, giảm thiểu chi phí và thời gian thực hiện các công tác an toàn. Tuy nhiên, khi số lượng các khu vực cần quan sát tăng lên hoặc những người chịu trách nhiệm quan sát không tập trung vào nhiệm vụ thì việc giám sát thông qua màn hình dễ xảy ra sai sót. Việc tích hợp công nghệ AI vào các hệ thống giám sát sẽ là sẽ tăng thêm độ tin cậy cho công tác đảm bảo an toàn, giảm thiểu những sai sót không đáng có.

1.2 Mục tiêu nghiên cứu

Mục tiêu của luận văn là xây dựng, đánh giá một hệ thống nhận diện việc sử dụng thiết bị bảo hộ cá nhân của người lao động trong công trường. Khi phát hiện ra các trường hợp không sử dụng các trang thiết bị bảo hộ thì hệ thống sẽ đưa ra cảnh báo.

1.3 Phạm vi nghiên cứu

Phạm vi của luận văn là tiến hành nhận dạng trên các video trích xuất từ camera. Mô hình nhận diện được huấn luyện sử dụng framework được xây dựng sẵn. Tập dữ liệu sử dụng để huấn luyện và đánh giá mô hình nhận diện được thu thập và dán nhãn bởi người làm luận văn. Một phần hình ảnh trong tập dữ liệu này có nguồn gốc từ các tập dữ liệu khác nhưng không sử dụng lại các nhãn của các tập dữ liệu đó. Các thiết bị bảo hộ cá nhân được tích hợp trong hệ thống gồm: mũ cứng, áo bảo hộ và khẩu trang.

1.4 Phương pháp nghiên cứu

Các giai đoạn trong quá trình nghiên cứu và hoàn thiện luận văn:

1. Tìm hiểu các mô hình nhận diện đang được nghiên cứu và sử dụng.
2. Chọn mô hình phù hợp, tìm hiểu lý thuyết.
3. Xây dựng tập dữ liệu phù hợp cho bài toán đặt ra.

4. Huấn luyện mô hình nhận diện.
5. Đánh giá các tham số hiệu năng của mô hình nhận diện đã xây dựng.
6. Xây dựng hệ thống sử dụng mô hình nhận diện để đưa ra cảnh báo.
7. Đánh giá khả năng hoạt động của hệ thống.

Trong đó các giai đoạn 3, 4 và 5 được thực hiện luân phiên và nhiều lần để cải thiện hiệu năng của mô hình nhận diện. YOLOv3 được chọn để làm mô hình nhận diện vì đây là một trong số các bộ nhận diện có hiệu năng cao trong thời gian thực, đã được sử dụng và đánh giá trên nhiều tập dữ liệu khác nhau.

1.5 Cấu trúc luận văn

Luận văn này bao gồm 5 chương. Chương 1 là chương mở đầu, giới thiệu bao quát về vấn đề, mục tiêu, phạm vi và phương pháp nghiên cứu luận văn. Chương 2 sẽ cung cấp những lý thuyết về các khái niệm được sử dụng trong luận văn. Chương 3 cho người đọc biết về cách tập dữ liệu được xây dựng, cách mô hình YOLOv3 được huấn luyện sử dụng framework darknet và cách xây dựng hệ thống sử dụng mô hình nhận diện. Chương 4 sẽ chứa những thông số đánh giá hiệu năng của bộ nhận diện và hệ thống trong các trường hợp khác nhau. Cuối cùng, trong chương 5 sẽ là các nhận xét về các kết quả đạt được và kết luận.

Chương 2

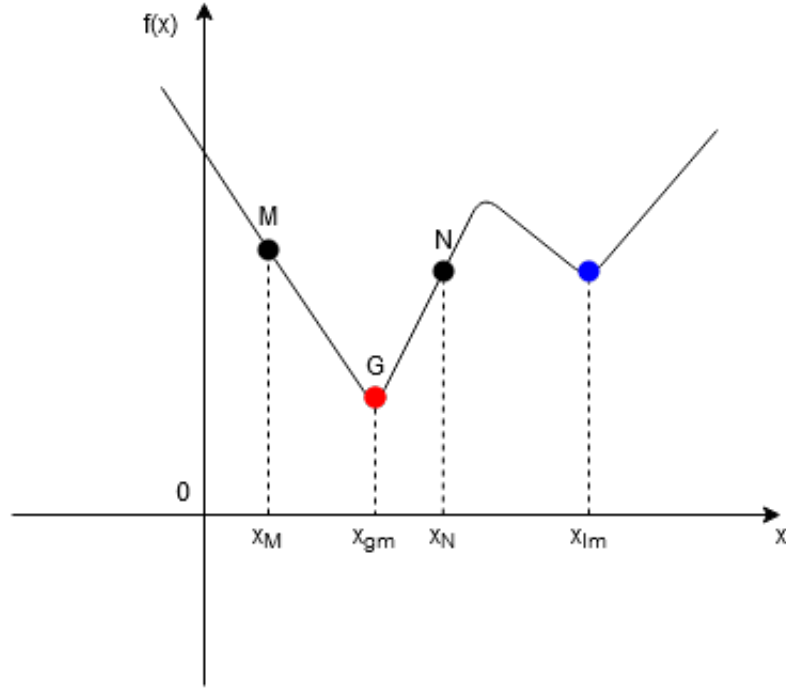
Cơ sở lý thuyết

2.1 Gradient Descent

Phần lớn các mô hình máy học được xây dựng dựa trên việc tối ưu hóa hàm mất mát hay nói cách khác là tìm cực tiểu của một hàm số biết trước. Đối với các hàm số đơn giản, việc xác định các điểm cực tiểu có thể được giải quyết thông qua việc tính toán đạo hàm cấp 1 và cấp 2. Tuy nhiên hàm mất mát của các mô hình máy học hay học sâu thường có số chiều lớn và đạo hàm phức tạp, do đó khó có thể áp dụng các phương pháp truyền thống để tìm các giá trị cực tiểu. Thay vào đó các mô hình này sử dụng giải thuật Gradient Descent để tìm các điểm cực tiểu của hàm mất mát.

Một hàm số có thể có nhiều điểm cực tiểu địa phương (local minimum) và cực tiểu toàn cục (global minimum). Ta có thể thấy trên hình 2.1 là đồ thị của một hàm số đơn biến, điểm màu xanh là cực tiểu địa phương, điểm màu đỏ là cực tiểu toàn cục. Giả sử ta có hai điểm M tại x_M và N tại x_N trên đồ thị hình 2.1, ta gọi điểm cực tiểu toàn cục là G . Lúc này ta muốn đưa điểm M và N về xấp xỉ hoặc trùng với vị trí của G bằng giải thuật Gradient Descent. Ta nhận thấy M nằm bên trái G và $f'(x_M) < 0$, nếu M muốn di chuyển về phía G thì $x_{M_{k+1}} = x_{M_k} + \delta$ tại bước thứ $k + 1$. Ngược lại nếu ta muốn N có $f'(x_N) > 0$ tiến về phía G tại bước tính toán thứ $k + 1$ thì $x_{N_{k+1}} = x_{N_k} - \delta$. Như vậy để một điểm bất kỳ $(x, f(x))$ lân cận G trên đồ thị tiến về G thì vị trí của điểm đó phải được cập nhật sau mỗi bước tính toán bằng cách cộng với một lượng δ với $sign(\delta) = -sign(f'(x))$. Trong thực tế, công thức được sử dụng có dạng:

$$x_{k+1} = x_k - \mu f'(x_k) \quad (2.1)$$



Hình 2.1: Cực tiểu địa phương (màu xanh) và cực tiểu toàn cục (màu đỏ)

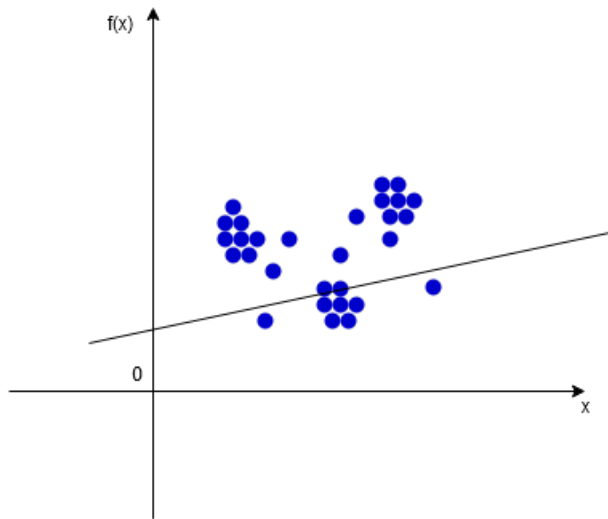
Với μ là *tốc độ học*, $\mu \in \mathbb{R}$, $\mu > 0$. Nếu ta chọn μ lớn thì ta sẽ cần ít số bước tính toán hơn để đến gần vị trí cực tiểu mong muốn nhưng trong nhiều trường hợp độ sai lệch giữa vị trí của điểm tính toán được sau cùng và vị trí của điểm cực tiểu sẽ tương đối cao. Ngược lại, nếu ta chọn μ nhỏ thì ta sẽ cần nhiều hơn số bước tính toán, bù lại khoảng cách giữa vị trí điểm tính toán được sau cùng và vị trí điểm cực tiểu sẽ có thể rất nhỏ.

Việc áp dụng giả thuật Gradient Descent lên làm đa biến là một sự mở rộng của ví dụ hàm đơn biến ở trên. Cho hàm số $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $n \in \mathbb{N}^*$, ta cần tìm cực tiểu cho $f(X)$ với $X = (x_0 \dots x_{n-1})$, $n \in \mathbb{N}^*$ từ một điểm khởi đầu X_0 bằng giải thuật Gradient Descent. Công thức để tính toán cho mỗi bước là:

$$X_{k+1} = X_k - \mu \nabla_X f(X_k) \quad (2.2)$$

2.1.1 Batch Gradient Descent

Giải thuật Batch Gradient Descent sử dụng tất cả các điểm đầu vào để cập nhật lại vector trọng số tại mỗi bước. Giả sử ta cần tối ưu hàm mất mát của một bài toán hồi quy tuyến tính gồm 30 điểm đầu vào với mỗi điểm gồm 2 tham số $(x, f(x))$ ở hình 2.2. Để tìm gradient cho mỗi điểm ta cần thực hiện 2 phép toán theo toán tử ∇ , đồng thời ta phải tìm gradient cho cả 30 điểm tại mỗi bước lặp và lấy trung bình của các kết quả này để cập nhật trọng số. Tổng số phép toán mà ta phải thực hiện cho tại mỗi bước là $30 \times 2 = 60$. Con số này sẽ tăng lên gấp nhiều lần đối với các bài toán thực tế khi số điểm và số tham số là vài triệu hoặc vài tỷ. Nói cách khác thuật toán này không hiệu quả về mặt tính toán với các bài toán máy học với dữ liệu lớn và phải cập nhật liên tục. Ngoài ra sau khi đã tìm được nghiệm tối

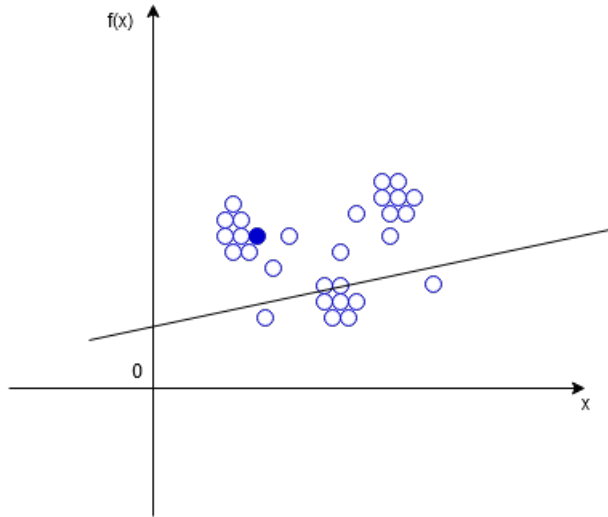


Hình 2.2: Batch Gradient Descent với bài toán hồi quy tuyến tính. Toàn bộ số điểm đầu vào đều được dùng để cập nhật các vector trọng số (a, b) cho đường hồi quy tại mỗi bước, với a là độ dốc và b độ sai lệch.

ưu của bài toán. Nếu ta thêm một điểm đầu vào mới vào tập dữ liệu cũ thì việc tính toán phải thực hiện lại từ đầu với toàn bộ điểm đầu vào bao gồm tập điểm đầu vào cũ và điểm mới thêm vào.

2.1.2 Stochastic Gradient Descent

Khác với Batch Gradient Descent giải thuật Stochastic Gradient Descent chỉ dùng gradient của một điểm ngẫu nhiên để cập nhật lại vector trọng số tại mỗi bước. Sau khi đi qua hết tất cả các điểm của tập đầu vào, thứ tự các điểm sẽ được xáo trộn và giải thuật lại tiếp tục với từng điểm. Mỗi một lần giải thuật Stochastic Gradient Descent tính toán xong với một điểm được gọi là một *iteration* còn với toàn bộ tập điểm thì gọi là một *epoch*. Cũng bài toán hồi quy tuyến tính ở trên nhưng với giải thuật Stochastic Gradient Descent (hình 2.3), ta có thể thấy số iteration mà giải thuật Stochastic Gradient Descent phải thực hiện trong một epoch là 30. Số phép tính của một lần tính toán là 2.



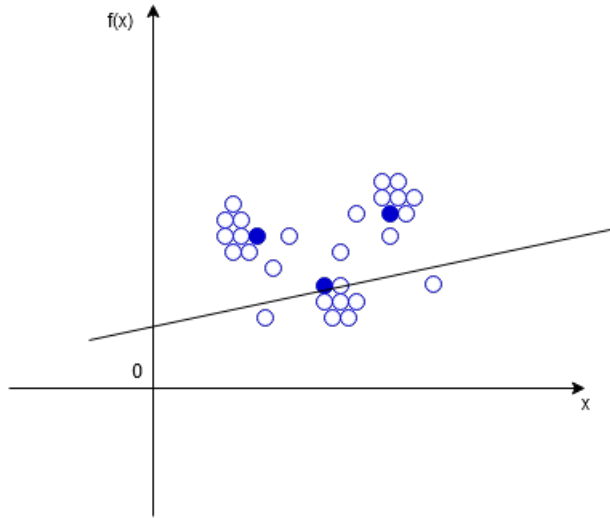
Hình 2.3: Stochastic Gradient Descent với bài toán hồi quy tuyến tính. Một điểm đầu vào được chọn ngẫu nhiên để cập nhật các vector trọng số (a, b) cho đường hồi quy tại mỗi iteration, với a là độ dốc và b độ sai lệch.

Do gradient của 1 điểm chỉ là xấp xỉ gần đúng của trung bình gradient của cả tập điểm nên việc cập nhật tại mỗi iteration sẽ có sai số nhất định, đồng thời các giá trị gradient tính toán được có thể có sự dao động lớn do tập điểm đầu vào thường bị tác động bởi nhiễu. Trên thực tế thì kết quả của giải thuật này có mức độ tối ưu khá tốt và hiệu quả tính toán cao. Sau khi đã hoàn thành tính toán trên tập dữ liệu cũ, nếu như có những điểm mới

được thêm vào thì ta chỉ cần chạy giải thuật với các điểm mới mà không cần phải chạy lại giải thuật với toàn bộ các điểm như Batch Gradient Descent.

2.1.3 Mini-batch Gradient Descent

Mini-batch Gradient Descent là sự kết hợp của Batch Gradient Descent và Stochastic Gradient Descent. Một mini-batch sẽ có n điểm với $1 < n \leq N$, N là tổng số điểm của tập dữ liệu đầu vào. Việc chia tập điểm ban đầu thành các batch sẽ được thực hiện một cách ngẫu nhiên. Mỗi một lần giải thuật xử lý xong một batch sẽ là một iteration và sau khi tất cả các batch được xử lý thì sẽ là một epoch. Như vậy $no_batch = \frac{N}{n}$. Phương pháp này cho kết quả gần với Batch Gradient Descent nhưng không dùng nhiều tài nguyên tính toán như Batch Gradient Descent và không cần phải lặp lại nhiều lần như Stochastic Gradient Descent.



Hình 2.4: Mini-batch Gradient Descent với bài toán hồi quy tuyến tính. Một batch sẽ gồm ba điểm đầu vào được chọn ngẫu nhiên để cập nhật các vector trọng số (a, b) cho đường hồi quy tại mỗi iteration, với a là độ dốc và b độ sai lệch. Một epoch sẽ gồm mười batch.

2.1.4 Điều kiện dừng của giải thuật

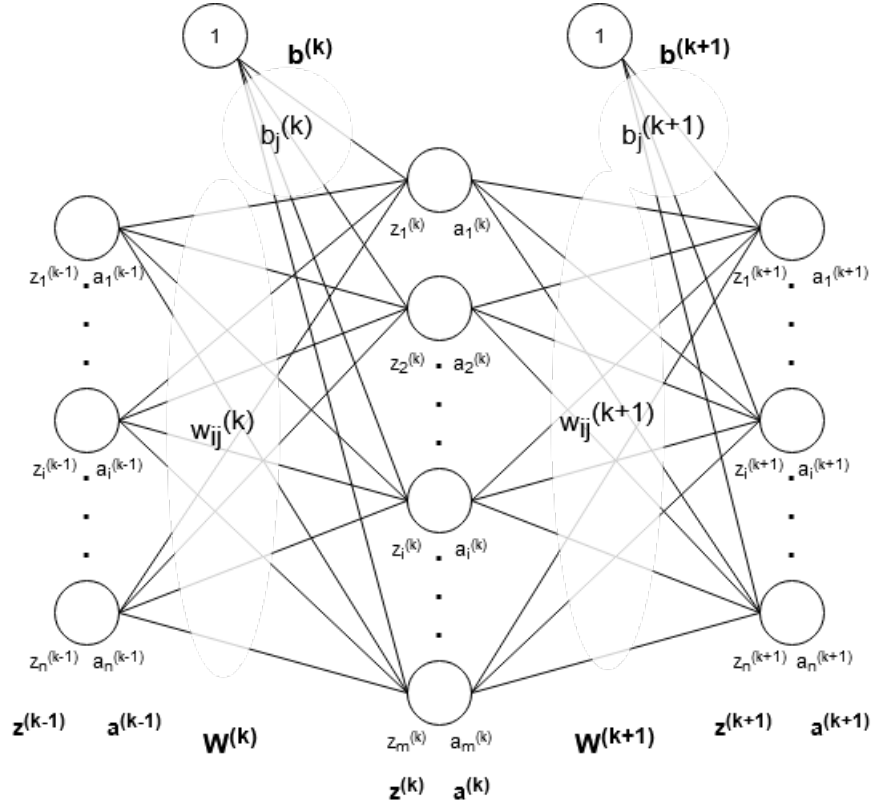
Ta đã biết các giải thuật Gradient Descent sẽ cần phải thực hiện rất nhiều vòng lặp tính toán để có thể hội tụ. Tuy nhiên rất khó để nói được khi nào có thể dừng được giải thuật. Trong thực tế có nhiều cách khác nhau được dùng để chọn số bước tính toán:

1. Chọn một số lượng vòng lặp nhất định dựa vào một số tiêu chí như số lượng dữ liệu đầu vào. Cách làm này có thể cho kết quả không đủ tốt, có thể nghiệm tối ưu nằm ở các bước trước hoặc sau điểm kết thúc.
2. Kiểm tra sự thay đổi của hàm mất mát giữa hai lần cập nhật liên tiếp, nếu sự sai lệch đạt tới ngưỡng đủ nhỏ thì ngưng giải thuật. Tuy nhiên nếu trên đồ thị của hàm mất mát có một vùng bằng phẳng nhưng không phải là cực tiểu thì giải thuật sẽ dừng tại điểm này mà không đạt được cực tiểu.
3. Kiểm tra sự thay đổi của gradient giữa hai lần cập nhật liên tiếp, nếu sự sai lệch đạt tới ngưỡng đủ nhỏ thì ngưng giải thuật. Nhược điểm của phương pháp này là việc tính gradient của các hàm phức tạp khó có thể thực hiện được.
4. Kiểm tra kết quả của giải thuật để ngừng việc lặp. Việc này cần người thực hiện việc huấn luyện mô hình phải thường xuyên kiểm tra các tham số hiệu năng của giải thuật lên một tập dữ liệu kiểm tra - *validation set* để xem tại thời điểm nào giải thuật có hiệu năng tốt nhất.

2.2 Backpropagation

Xét một mô hình mạng neuron (hình 2.5) Quá trình dữ liệu được đưa vào lớp đầu tiên cho đến khi có kết quả ở lớp sau cùng được gọi là quá trình *feed-forward*.

$$\begin{aligned}a^{(0)} &= x \\z^{(k)} &= \mathbf{W}^{(k)T} a^{(k-1)} + \mathbf{b}^k, k = 1, 2, \dots, N \\a^{(l)} &= f^{(k)}(z^{(k)}), k = 1, 2, \dots, N \\\hat{y} &= a^{(N)}\end{aligned}$$



Hình 2.5: Mô hình mạng neuron đơn giản.

Ta thấy với mô hình này, hàm mất mát $J(\mathbf{W}, \mathbf{b}, \mathbf{X}, \mathbf{Y})$ sẽ phụ thuộc vào tập các ma trận trọng số \mathbf{W} và tập các vector bias của mỗi lớp \mathbf{b} . Việc tính gradient của hàm mất mát phụ thuộc vào việc tính các đạo hàm riêng $\frac{\partial J}{\partial \mathbf{W}^{(k)}}$; $\frac{\partial J}{\partial \mathbf{b}^{(k)}}$, $\forall k = 1, 2, \dots, N$. Đối với bài toán hồi quy tuyến tính thì hàm mất mát là hàm trung bình bình phương sai số (Mean Square Error - MSE), lúc này

$$J(\mathbf{W}, \mathbf{b}, \mathbf{X}, \mathbf{Y}) = \frac{1}{M} \sum_{i=1}^{(M)} \|y_i - \hat{y}_i\|_2^2 = \frac{1}{M} \sum_{i=1}^{(M)} \|y_i - a_i^{(N)}\|_2^2 \quad (2.3)$$

Với N là số điểm trong tập điểm đầu vào. Ta nhận thấy để tìm các đạo hàm riêng của J với \mathbf{W} và \mathbf{b} trong trường hợp này là rất khó vì phương trình của J không phụ thuộc trực tiếp vào \mathbf{W} và \mathbf{b} . Để có thể hiện thực các giải thuật thuộc họ Gradient Descent thì phương pháp thường được sử dụng là

Backpropagation. Phương pháp này sẽ cập nhật các trọng số theo chiều từ layer cuối cùng đến layer đầu tiên. Đầu tiên giải thuật sẽ tính đạo hàm của hàm mất mát theo ma trận trọng số của lớp cuối cùng.

$$\begin{aligned}\frac{\partial J}{\partial w_{ij}^{(N)}} &= \frac{\partial J}{\partial z_j^{(N)}} \cdot \frac{\partial z_j^{(N)}}{\partial w_{ij}^{(N)}} \\ &= e_j^{(N)} \frac{\partial \left(w_{ij}^{(N)T} a^{(N-1)} + b_j^{(N)} \right)}{\partial w_{ij}^{(N)}} \\ &= e_j^{(N)} a_i^{(N-1)}\end{aligned}$$

Với $e_j^{(N)} = \frac{\partial J}{\partial z_j^{(N)}}$ có thể tính được tương đối dễ dàng. Tương tự ta có đạo hàm riêng của J với bias ở lớp cuối cùng.

$$\begin{aligned}\frac{\partial J}{\partial b_j^{(N)}} &= \frac{\partial J}{\partial z_j^{(N)}} \cdot \frac{\partial z_j^{(N)}}{\partial b_j^{(N)}} \\ &= e_j^{(N)}\end{aligned}$$

Các công thức trên cũng đúng với một lớp bất kỳ trong mạng neuron. Ta lấy mô hình hai lớp liên tiếp của một mạng neuron ở hình 2.5 để đưa ra công thức tổng quát như sau:

$$\begin{aligned}\frac{\partial J}{\partial w_{ij}^{(k)}} &= \frac{\partial J}{\partial z_j^{(k)}} \cdot \frac{\partial z_j^{(k)}}{\partial w_{ij}^{(k)}} \\ &= e_j^{(k)} \frac{\partial \left(w_{ij}^{(k)T} a^{(k-1)} + b_j^{(k)} \right)}{\partial w_{ij}^{(k)}} \\ &= e_j^{(k)} a_i^{(k-1)}\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial b_j^{(k)}} &= \frac{\partial J}{\partial z_j^{(k)}} \cdot \frac{\partial z_j^{(k)}}{\partial b_j^{(k)}} \\ &= e_j^{(k)}\end{aligned}$$

Ta sẽ tính $e_j^{(k)}$ như sau:

$$\begin{aligned}
e_j^{(k)} &= \frac{\partial J}{\partial z_j^{(k)}} = \frac{\partial J}{\partial a_j^{(k)}} \cdot \frac{\partial a_j^{(k)}}{\partial z_j^{(k)}} \\
&= \left(\sum_{l=1}^{d^{(k+1)}} \frac{\partial J}{\partial z_l^{(k+1)}} \cdot \frac{\partial z_l^{(k+1)}}{\partial a_j^{(k)}} \right) f^{(k)'}(z_j^{(k)}) \\
&= \left(\sum_{l=1}^{d^{(k+1)}} e_l^{(k+1)} \cdot w_{jl}^{(k+1)} \right) f^{(k)'}(z_j^{(k)})
\end{aligned}$$

Ta có $f : \mathbb{R} \rightarrow [0, 1]$ là hàm kích (activation function) hay còn gọi là hàm bao tại một node trong mạng neuron, $a_j^k = f(z_j^k)$, do đó ta có đạo hàm riêng của a_j^k theo z_j^k chính là đạo hàm của f . Ngoài ra do a_j^k trực tiếp tham gia vào việc tính các $z_l^{k+1}, l = 1, 2, \dots, d^{(k+1)}$ nên $\frac{\partial J}{\partial a_j^k}$ có thể tách ra thành tổng của tích các đạo hàm riêng như dòng thứ hai. Tương tự như vậy ta có thể tính

$$\frac{\partial J}{\partial b_j^{(k)}} = e_j^{(k)} \quad (2.4)$$

Việc tính e_j^k sẽ phụ thuộc vào kết quả của e_j^{k+1} do đó phương pháp này được gọi là Backpropagation. Các bước để thực hiện giải thuật Backpropagation cho một mạng neuron nhân tạo gồm:

1. Feedforward: Với mỗi giá trị đầu vào của x , tính giá trị đầu ra của mạng neuron, đồng thời lưu lại các kết quả $\mathbf{a}^{(k)}$ tại mỗi lớp.
2. Với mỗi node thứ j ở lớp ngoài cùng tính

$$e_j^{(N)} = \frac{\partial J}{\partial z_j^{(N)}} \quad (2.5)$$

3. Từ đó suy ra:

$$\begin{aligned}
\frac{\partial J}{\partial w_{ij}^{(N)}} &= a_i^{(N-1)} e_j^{(N)} \\
\frac{\partial J}{\partial b_j^{(N)}} &= e_j^{(N)}
\end{aligned}$$

4. Với $k = N - 1, N - 2, \dots, 1$ tìm $e_j^{(k)}$

$$e_j^{(k)} = \left(\sum_{l=1}^{d^{(k+1)}} e_l^{(k+1)} \cdot w_{jl}^{(k+1)} \right) f^{(k)'} \left(z_j^{(k)} \right) \quad (2.6)$$

5. Cập nhật đạo hàm cho từng trọng số và bias:

$$\begin{aligned} \frac{\partial J}{\partial w_{ij}^{(k)}} &= a_i^{(k-1)} e_j^{(k)} \\ \frac{\partial J}{\partial b_j^{(k)}} &= e_j^{(k)} \end{aligned}$$

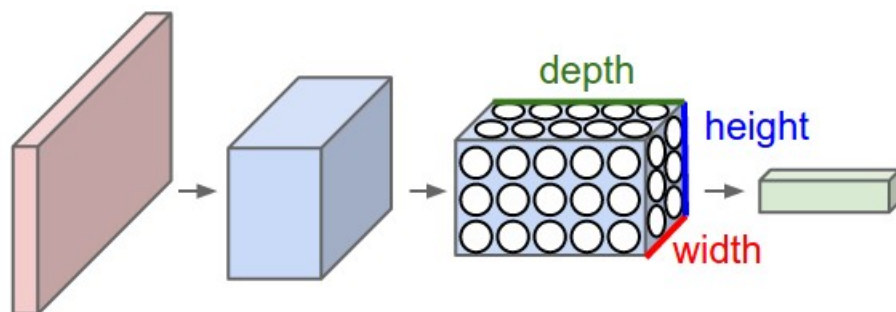
2.3 Mạng neuron tích chập

Mạng neuron tích chập (tiếng Anh: Convolutional Neural Network - CNN) là một loại mạng neuron dùng riêng cho các bài toán về hình ảnh. Bên trong mạng neuron tích chập vẫn là các neuron có các trọng số và bias có thể cập nhật được để học các đặc trưng của hình ảnh.

Các lớp của mạng neuron tích chập được bố trí theo ba chiều: chiều rộng (tiếng Anh: width), chiều cao (tiếng Anh: height), chiều sâu (tiếng Anh: depth). Chiều sâu ở đây muốn nói tới chiều sâu của miền các neuron kích hoạt (tiếng Anh: activation volume) chứ không phải là chiều sâu của cả mạng neuron. Các neuron ở lớp sau sẽ chỉ được kết nối với một phần nhỏ các neuron ở lớp trước chứ không phải là toàn bộ như trong các mạng neuron thông thường. Ta lấy ví dụ mạng CIFAR-10 (hình 2.6), miền các neuron kích hoạt ở mạng neuron này có kích thước các chiều là $32 \times 32 \times 3$ (*rộng* \times *cao* \times *sâu*). Lớp cuối cùng của CIFAR-10 sẽ có kích thước các chiều là $1 \times 10 \times 10$ ứng với vector điểm cho các nhãn cần được phân loại (tiếng Anh: class scores). Một mạng neuron tích chập thông thường sẽ được cấu tạo từ ba loại lớp neuron: lớp tích chập (tiếng Anh: convolutional layer), lớp pooling (tiếng Anh: pooling layer) và lớp đầy đủ kết nối (tiếng Anh: fully-connected layer).

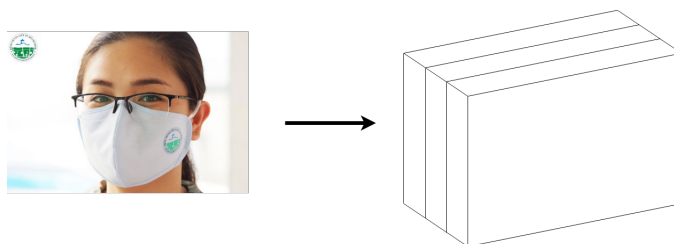
2.3.1 Lớp tích chập

Trong lớp này đầu vào lớp đầu tiên sẽ là một ảnh màu có ba kênh màu: đỏ, xanh lá, xanh dương (hình 2.7). Đầu ra của các lớp trước sẽ là đầu vào



Hình 2.6: Mô hình mạng neuron tích chập đơn giản. Lớp nhận hình ảnh vào màu đỏ là một lớp có cấu trúc ba chiều với chiều rộng và chiều cao là chiều rộng và chiều cao của hình ảnh đầu vào, chiều sâu bằng ba ứng với ba kênh màu đỏ, xanh lá và xanh dương. Các lớp của mạng neuron tích chập sẽ chuyển đổi một nhóm các ma trận thành một nhóm các ma trận khác. Lớp ngoài cùng là lớp phân loại, có kích thước các chiều tương ứng với một vector.

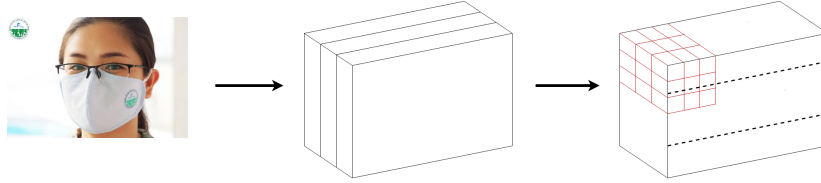
của các lớp sau. Các tensor trong mạng tích chập được gọi là các tensor. Sau



Hình 2.7: Hình ảnh đầu vào gồm ba kênh màu được mô hình hóa thành tensor với chiều cao và chiều rộng là chiều cao và chiều rộng của ảnh, chiều sâu là ba.

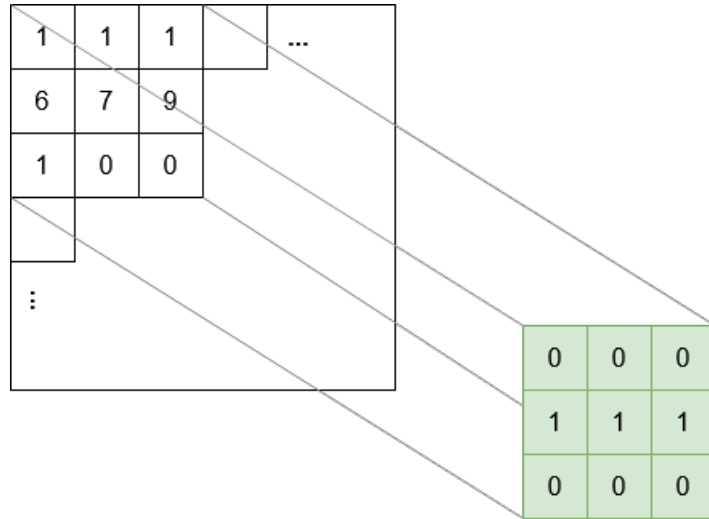
đó một bộ lọc có kích thước $m \times n \times 3$ (tiếng Anh: kernel) sẽ được trượt qua tensor của ảnh đầu vào. Ở mỗi kênh màu, lớp tương ứng của kernel sẽ hoạt động như một cửa sổ trượt (tiếng Anh: sliding window). Nhắc lại một chút về phép toán của cửa sổ trượt trên ảnh trắng đen. Giả sử ta có một cửa sổ trượt có kích thước 3×3 đang quét qua một hình trắng đen (hình 2.9), tại vị trí như trên hình việc tính toán giá trị đầu ra được thực hiện như sau

$$1 \times 0 + 1 \times 0 + 1 \times 0 + 6 \times 1 + 7 \times 1 + 9 \times 1 + 1 \times 0 + 0 \times 0 + 0 \times 0 = 22 \quad (2.7)$$



Hình 2.8: Hình ảnh sau khi được đưa qua đầu vào và chuyển đổi thành dữ liệu ba chiều sẽ được đưa vào lớp convolution đầu tiên. Một kernel có kích thước $3 \times 3 \times 3$ (góc trên bên trái của mô hình ngoài cùng bên phải) được trượt qua hình đầu vào.

Ngoài ra, ta còn hai khái niệm cần nhắc tới là stride và padding.



Hình 2.9: Ví dụ về phép toán của sổ trượt với kích thước 3×3 .

- Với stride bằng một thì cửa sổ trượt sẽ di chuyển tuần tự qua tất cả các ô của ma trận. Tổng quát với $stride = k$ thì các điểm ảnh được cửa sổ trượt đi qua của một ma trận có kích thước $m \times n$ sẽ là $x_{1+i \times k, 1+j \times k}$ với $i, j \in \mathbb{N}; 1 + i \times k \leq m; 1 + j \times k \leq n$.
- Đối với các điểm ảnh ở gần biên, nếu như trong vùng cửa sổ không có những chỗ không tồn tại giá trị điểm ảnh thì các phương pháp chèn giá trị (tiếng Anh: padding) sẽ được sử dụng để thay làm các giá trị tính

toán. Một trong các cách padding phổ biến là dùng các giá trị bằng không (tiếng Anh: zero padding).

| | | | | |
|---|---|---|----|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 2 | 3 | -2 | 0 |
| 0 | 9 | 1 | 0 | 0 |
| 0 | 1 | 2 | 3 | 0 |
| 0 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|----|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 2 | 3 | -2 | 0 | 0 |
| 0 | 0 | 9 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 2 | 3 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Hình 2.10: Bên trái, ma trận 3×3 được zero padding với $padding = 1$. Bên phải, ma trận 3×3 được zero padding với $padding = 2$

Như vậy, nếu đầu vào của phép tính tích chập là ma trận X có kích thước $m \times n$ với cửa sổ trượt có kích thước $k \times k$, $stride = s$, $padding = p$ thì đầu ra sẽ là một ma trận Y có kích thước $\left(\frac{m-k+2p}{s} + 1\right) \times \left(\frac{n-k+2p}{s} + 1\right)$

Việc tính toán tại mỗi kênh màu của hình khi kernel đi qua cũng gần tương tự với cửa sổ trượt. Kết quả phép toán của ba kênh màu và một bias sẽ được cộng lại và đưa vào ma trận kết quả. Giả sử ta có một kernel có kích thước $3 \times 3 \times 3$ như hình 2.11. Dữ liệu một ảnh đầu vào gồm ba kênh màu,

| | | |
|----|----|----|
| -1 | 0 | 1 |
| 0 | 1 | -1 |
| 1 | -1 | -1 |

| | | |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 0 | 0 |

| | | |
|---|---|---|
| 0 | 1 | 0 |
| 0 | 1 | 0 |
| 0 | 1 | 0 |

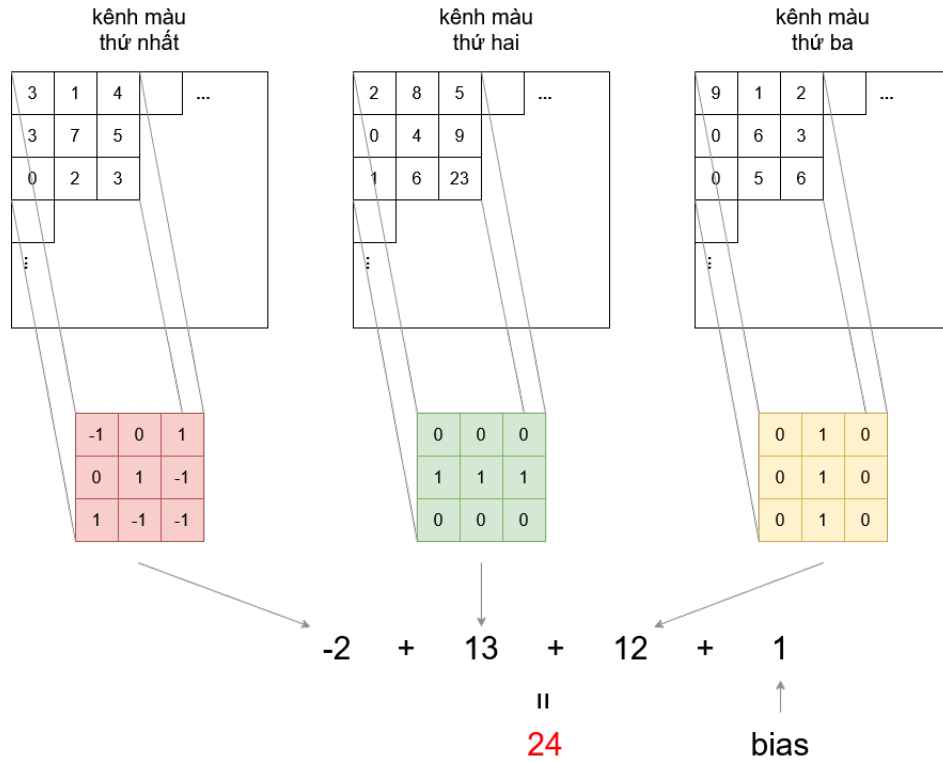
kênh kernel
thứ nhất

kênh kernel
thứ hai

kênh kernel
thứ ba

Hình 2.11: Ví dụ về một kernel có kích thước $3 \times 3 \times 3$.

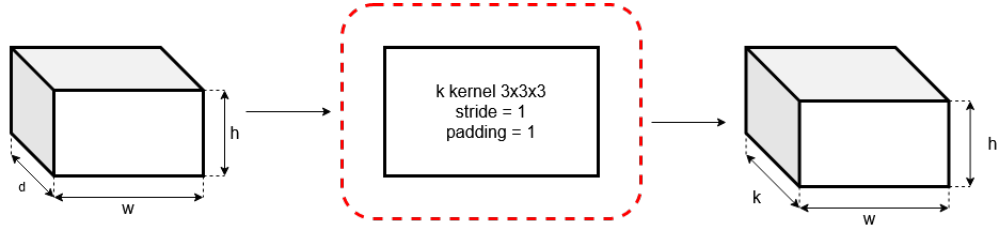
khi đi qua lớp tích chập đầu tiên sẽ được tính toán như hình 2.12 Sau khi



Hình 2.12: Ví dụ về phép toán của một kernel lên một vị trí của ảnh trong lớp tích chập.

kernel đã quét qua hết các điểm ảnh mong muốn thì kết quả nhận được sẽ là một ma trận. Mỗi một kernel khác nhau sẽ trích xuất ra được một đặc trưng khác nhau của ảnh. Do đó một lớp tích chập sẽ có nhiều kernel để lấy các đặc trưng khác nhau. Lúc này đầu ra sẽ là một tensor gồm nhiều ma trận. Nếu như có k kernel được dùng tại một lớp tích chập thì đầu ra sẽ có chiều sâu bằng k , chiều rộng và chiều cao sẽ bằng chiều rộng và chiều cao của ảnh đầu vào. Đầu ra của lớp tích chập trước sẽ là đầu vào của lớp tích chập sau. Tổng quát hóa, với một lớp tích chập với K kernel có kích thước $N \times N \times D$ (với D là chiều sâu của đầu vào và là số lẻ), $stride = S$, $padding = P$. Đầu vào là một tensor có kích thước $H \times W \times D$ thì kích thước của tensor đầu ra sẽ là $\left(\frac{H-F+2P}{S} + 1\right) \times \left(\frac{W-F+2P}{S} + 1\right) \times K$

Đầu ra của lớp tích chập sẽ đi qua hàm kích hoạt trước khi được đưa vào lớp tích chập tiếp theo. Mỗi kernel với kích thước $N \times N \times D$ sẽ có một hệ số bias tương ứng với tổng số các tham số của một kernel là $N \times N \times D + 1$,



Hình 2.13: Một lớp tích chập có k kernel với kích thước $3 \times 3 \times 3$, $stride = 1$, $padding = 1$. Đầu vào là một tensor có kích thước $h \times w \times d$ đầu ra của phép tích chập lên tensor này khi khối tích chập có các thông số ở trên là một tensor có kích thước $h \times w \times k$

với K kernel thì số tham số sẽ là $K \times (N \times N \times D + 1)$.

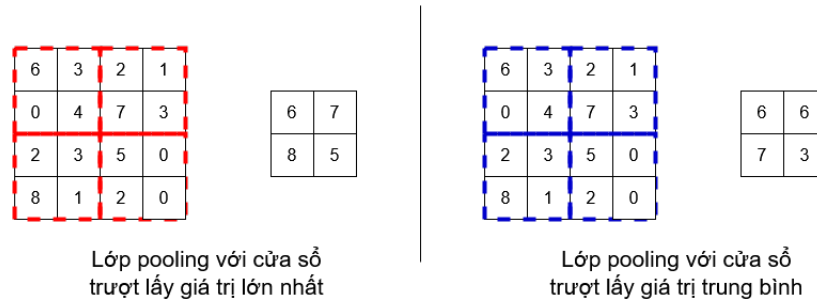
2.3.2 Lớp pooling

Việc tính toán với toàn bộ dữ liệu đầu vào của ảnh có độ phân giải lớn và kích thước lớn trên mạng neuron tích chập thường không hiệu quả về mặt tính toán do sẽ có nhiều điểm ảnh miêu tả cùng một đặc trưng. Do đó lớp pooling được dùng ở giữa các lớp tích chập để giảm kích thước của các tensor nhưng vẫn không làm mất đi các đặc trưng của dữ liệu.

Cho một lớp pooling có kích thước cửa sổ trượt là $N \times N$, đầu vào là một tensor có kích thước $H \times W \times D$. Ta chia tensor này thành D ma trận $H \times W$. Với mỗi ma trận ta lần lượt trượt cửa sổ trượt của lớp pooling lên từng điểm ảnh. Trong vùng dữ liệu của cửa sổ trượt ta sẽ tìm giá trị lớn nhất hoặc trung bình của các giá trị để đưa vào ma trận mới. Một số mô hình mạng neuron tích chập sẽ dùng $stride > 1$ trong lớp tích chập để làm giảm kích thước dữ liệu thay vì dùng lớp pooling. Ngoài ra, trong thực tế lớp pooling thường được sử dụng với kích thước cửa sổ trượt 2×2 , $stride = 2$, $padding = 0$. Chiều cao và chiều rộng của tensor đầu ra sẽ giảm đi một nửa còn chiều sâu vẫn giữ nguyên.

2.3.3 Lớp đầy đủ kết nối

Hình ảnh sau khi qua các lớp tích chập và pooling thì đầu ra sẽ là một tensor chứa các đặc trưng mà mô hình trích xuất được. Tensor có kích thước $H \times W \times D$ tại lớp tích chập cuối cùng sẽ được chuyển thành một vector có

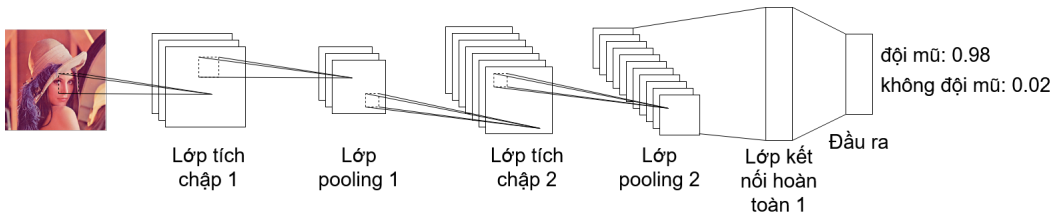


Hình 2.14: Bên trái, lớp pooling với cửa sổ trượt lấy giá trị lớn nhất với kích thước cửa sổ 2×2 , $stride = 1$, $padding = 0$. Bên phải, lớp pooling với cửa sổ trượt lấy giá trị trung bình với kích thước cửa sổ 2×2 , $stride = 2$, $padding = 0$.

chiều dài $H \times W \times D$. Sau đó vector này sẽ được đưa vào các lớp đầy đủ kết nối để đưa ra kết quả dự đoán cho ảnh.

2.3.4 Mô hình mạng neuron tích chập

Ảnh đầu vào \rightarrow [Lớp tích chập \rightarrow Lớp pooling] $\times n \rightarrow$ [Lớp liên kết hoàn toàn] $\times m \rightarrow$ Đầu ra, với $m, n \in \mathbb{N}^*$.



Hình 2.15: Mạng neuron tích chập gồm hai lớp tích chập và pooling, một lớp kết nối đầy đủ.

2.4 YOLOv3

YOLO - You Only Look Once là một trong những mô hình nhận diện thời gian thực hiện đại nhất và đang được sử dụng cho nhiều bài toán nhận diện, theo dõi khác nhau. Không như những mạng CNN hay R-CNN trước

đây, thay vì sử dụng phương pháp dự đoán trên từng miền (tiếng Anh: region proposal method) và cửa sổ trượt (tiếng Anh: sliding window) để phát hiện vật thể trong từng vùng nhỏ trong khung hình và tiến hành phân loại vật thể đó. YOLO sử dụng cả khung hình để nhận diện vật thể, các đặc trưng của vùng nền phía sau (tiếng Anh: background) cũng được dùng trong quá trình huấn luyện. Do vậy YOLO có thể nhận diện vật thể một cách nhanh chóng trong một khung hình với độ chính xác cao chỉ với một lần xử lý, đó cũng là lý do vì sao mô hình này được gọi là Bạn Chỉ Cần Nhìn Một Lần (tiếng Anh: You Only Look Once).

2.4.1 Unified Detection

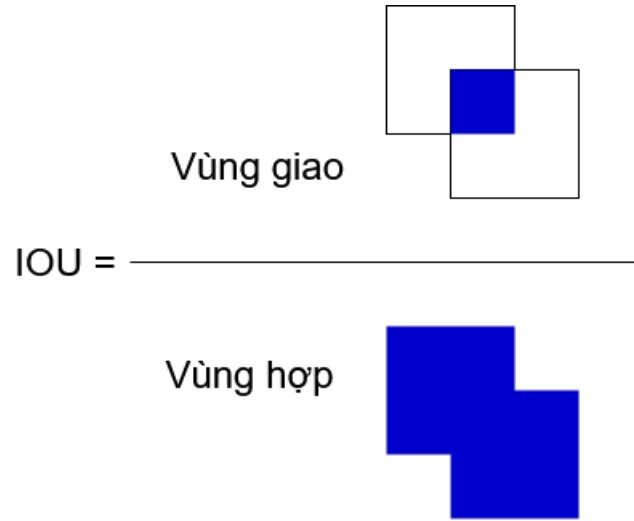
Hình ảnh đầu vào được chia thành một mạng lưới ô vuông (tiếng Anh: grid cell) có kích thước $S \times S$. Nếu như tâm của một vật thể nằm ở tâm của một ô thì ô đó sẽ chịu trách nhiệm trong việc nhận diện vật thể đó. Mỗi



Hình 2.16: Hình ảnh được chia thành mạng lưới ô vuông $S \times S$.

ô sẽ dự đoán B bounding box và độ tin cậy (tiếng Anh: confidence score) của từng bounding box. Gọi $Pr(Object)$ là xác suất của vật thể nằm trong một ô với $Pr(Object) \in \mathbb{R}, 0 \leq Pr(Object) \leq 1$. IOU - intersection over union là tỷ lệ giữa diện tích miền giao và diện tích miền hợp của bounding box dự đoán được và bounding box được tạo sẵn để huấn luyện (tiếng Anh: ground truth) hình 2.17, $IOU \in \mathbb{R}, 0 \leq IOU \leq 1$. Độ tin cậy sẽ được định nghĩa bằng $Pr(Object) \times IOU$, nếu như một ô không chứa vật thể thì $Pr(Object) = 0$ suy ra độ tin cậy sẽ bằng không, ngược lại nếu một ô chứa vật thể thì $Pr(Object) = 1$, lúc này độ tin cậy bằng IOU .

Mỗi một bounding box sẽ có năm tham số cần dự đoán: t_x, t_y, t_w, t_h và độ tin cậy. (t_x, t_y) là tọa độ tương đối của tâm bounding box với một ô, nếu gọi



Hình 2.17: Miêu tả việc tính toán IOU.

offset của một ô trong hình là (c_x, c_y) thì tọa độ của một bounding box so với hình sẽ là $(\sigma(t_x) + c_x, \sigma(t_y) + c_y)$. Nếu chiều dài và chiều rộng của bounding box cho trước là (p_w, p_h) chiều dài và chiều rộng của bounding box được sự đoán sẽ là $(p_w \times e^{t_w}, p_h \times e^{t_h})$.

$$b_x = \sigma(t_x) + c_x$$

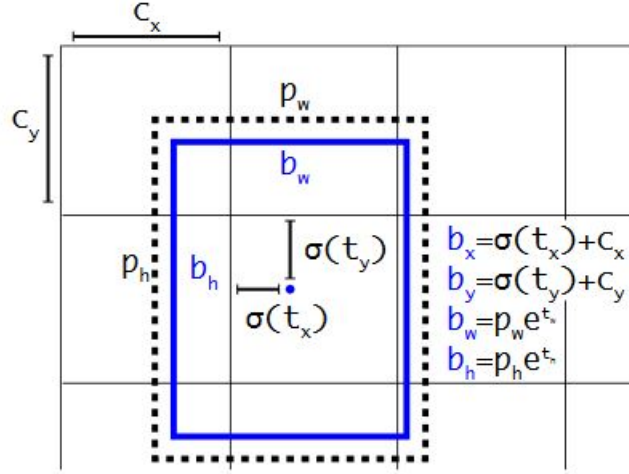
$$b_y = \sigma(t_y) + c_y$$

$$b_w = p_w \times e^{t_w}$$

$$b_h = p_h \times e^{t_h}$$

Trong quá trình huấn luyện, hàm mất mát tổng bình phương sai số được sử dụng cho các tọa độ của bounding box. Với một dự đoán t_* có ground truth là \hat{t}_* thì gradient sẽ là hiệu giữa ground truth và kết quả dự đoán được $\hat{t}_* - t_*$.

YOLOv3 sử dụng hàm hồi quy logistic để dự đoán xem một bounding box có chứa vật thể hay không. Nếu như bounding box dự đoán được giao với ground truth nhiều hơn các bounding box trước đó thì kết quả là 1. Nếu như một bounding box dự đoán được không phải là trường hợp có diện tích giao lớn nhất với ground truth nhưng vẫn có độ tin cậy lớn hơn một ngưỡng xác quyết thì dự đoán với bounding box này sẽ bị bỏ qua. Ngưỡng xác quyết được dùng là 0.5. Nếu như một bounding box không được đặt vào



Hình 2.18: Mô hình dự đoán bounding box của YOLO.

một ground truth thì sẽ không có các dự đoán về tọa độ và class mà chỉ có dự đoán về sự tồn tại của vật thể.

Độ tin cậy chính là IOU giữa bounding box dự đoán được và bounding box được tạo sẵn. Đối với những ô được dự đoán có vật thể, mô hình sẽ dự đoán thêm C xác suất của các class mà vật thể đó thuộc về $Pr(Class_i|Object)$ với $Pr(Class_i|Object) \in \mathbb{R}, 0 \leq Pr(Class_i|Object) \leq 1, i = 1, \dots, C$. Mỗi ô sẽ chỉ có một tập các giá trị $Pr(Class_i|Object)$ mà không liên quan tới số lượng bounding box B .

Khi tiến hành dự đoán, YOLO sẽ nhân các xác suất và độ tin cậy lại với nhau để được xác suất của một class trên một bounding box.

$$Pr(Class_i|Object) \times Pr(Object) \times IOU = Pr(Class_i) \times IOU \quad (2.8)$$

Hàm mất mát được sử dụng khi huấn luyện để dự đoán các class là hàm binary cross-entropy nhằm giúp mô hình có thể dự đoán đa lớp trong cùng một ô.

Sau cùng ta có hàm mất mát tổng của YOLOv3

$$\begin{aligned}
Loss = & \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbf{1}_{ij}^{obj} \left[(\hat{t}_{x_{ij}} - t_{x_{ij}})^2 + (\hat{t}_{y_{ij}} - t_{y_{ij}})^2 + (\hat{t}_{w_{ij}} - t_{w_{ij}})^2 + (\hat{t}_{h_{ij}} - t_{h_{ij}})^2 \right] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^B - [\hat{p}_{o_{ij}} \times \log(p_{o_{ij}}) + (1 - \hat{p}_{o_{ij}}) \times \log(1 - p_{o_{ij}})] \\
& + \sum_{i=0}^{S^2} \mathbf{1}_{ij}^{obj} \sum_{k=1}^C - [\hat{p}_i(c_k) \times \log(p_i(c_k)) + (1 - \hat{p}_i(c_k)) \times \log(1 - p_i(c_k))]
\end{aligned}$$

2.4.2 Kiến trúc mạng YOLOv3

Các đặc trưng từ ảnh được trích xuất theo ba tỷ lệ khác nhau bằng phương pháp giống như phương pháp được sử dụng trong mạng trích xuất đặc trưng dạng kim tự tháp (tiếng Anh: feature pyramid networks). Đầu ra của miền trích xuất đặc trưng là một tensor ba chiều chứa vị trí bounding box, xác suất tồn tại vật thể, xác suất các class, kích thước của tensor là $S \times S \times [3 \times (4 + 1 + C)]$, với $S \times S$ là số ô mà ảnh được chia thành, 4 là các giá trị dự đoán của bounding box (t_x, t_y, t_w, t_h), 1 là giá trị dự đoán sự tồn tại của vật thể trong ô, C là vector các giá trị dự đoán của các class.

Sau đó các ma trận đặc trưng từ hai lớp trước và upsample lên hai lần. Ngoài ra các ma trận đặc trưng từ các lớp đầu cũng được ghép lại với các lớp sau. Việc này giúp mô hình lấy được nhiều thông tin có ý nghĩa từ các ma trận đặc trưng được upsample và vẫn giữ được những đặc trưng nhỏ hơn từ các lớp đầu. Sau đó một vài lớp tích chập được sử dụng để kết hợp các ma trận đặc trưng và đưa ra tensor sau cùng chứa các dự đoán có kích thước gấp đôi.

Với tỷ lệ cuối, mô hình như ở trên được lặp đi lặp lại nhiều lần, điều này cho phép các bounding box ở tỷ lệ cuối có thể dùng các giá trị đã được tính toán từ các tỷ lệ trước cũng nhưng các đặc trưng có ý nghĩa.

Ban đầu sẽ có chín bounding box: $(10 \times 13), (16 \times 30), (33 \times 23), (30 \times 61), (62 \times 45), (59 \times 119), (116 \times 90), (156 \times 198), (373 \times 326)$. Các bounding box này sẽ được chia vào ba tỷ lệ bằng giải thuật k-means. Các bounding box này sẽ được dùng để dự đoán ở các tỷ lệ tương ứng. Ảnh đầu vào sẽ được downsample với các stride bằng 32, 16 và 8 cho từng tỷ lệ. Nhờ vậy YOLOv3 có thể dự đoán được các vật thể nhỏ rất tốt.

YOLOv3 sử dụng kiến trúc mạng kết hợp giữa YOLOv2, Darknet-19 và mạng residual. Các lớp tích chập có kích thước 3×3 , 1×1 và có thêm các liên kết tắt (tiếng Anh: short cut). Có tất cả 53 lớp tích chập được dùng nên kiến trúc này được gọi là Darknet-53. Kiến trúc này cho khả năng nhận diện

| | Type | Filters | Size | Output |
|----|---------------|---------|------------------|------------------|
| | Convolutional | 32 | 3×3 | 256×256 |
| | Convolutional | 64 | $3 \times 3 / 2$ | 128×128 |
| 1x | Convolutional | 32 | 1×1 | |
| | Convolutional | 64 | 3×3 | |
| | Residual | | | 128×128 |
| | Convolutional | 128 | $3 \times 3 / 2$ | 64×64 |
| 2x | Convolutional | 64 | 1×1 | |
| | Convolutional | 128 | 3×3 | |
| | Residual | | | 64×64 |
| | Convolutional | 256 | $3 \times 3 / 2$ | 32×32 |
| 8x | Convolutional | 128 | 1×1 | |
| | Convolutional | 256 | 3×3 | |
| | Residual | | | 32×32 |
| | Convolutional | 512 | $3 \times 3 / 2$ | 16×16 |
| 8x | Convolutional | 256 | 1×1 | |
| | Convolutional | 512 | 3×3 | |
| | Residual | | | 16×16 |
| | Convolutional | 1024 | $3 \times 3 / 2$ | 8×8 |
| 4x | Convolutional | 512 | 1×1 | |
| | Convolutional | 1024 | 3×3 | |
| | Residual | | | 8×8 |
| | Avgpool | | Global | |
| | Connected | | 1000 | |
| | Softmax | | | |

Hình 2.19: Kiến trúc mạng Darknet-53.

tốt hơn Darknet-19 và hiệu năng cao hơn ResNet-101 hoặc ResNet-152 bằng 2.1.

| <i>Backbone</i> | <i>Top-1</i> | <i>Top-5</i> | <i>Bn Ops</i> | <i>BFLOP/s</i> | <i>FPS</i> |
|-----------------|--------------|--------------|---------------|----------------|------------|
| Darknet-19 | 74.1 | 91.8 | 7.29 | 1246 | 171 |
| ResNet-101 | 77.1 | 93.7 | 19.7 | 1039 | 53 |
| ResNet-152 | 77.6 | 93.8 | 29.4 | 1090 | 37 |
| Darknet-53 | 77.2 | 93.8 | 18.7 | 1457 | 78 |

Bảng 2.1: So sánh hiệu năng của Darknet-53 với các mạng khác. Accuracy, Bn Ops - billions of operations, BFLOP/s - billion floating point operations per second, và FPS - frames per second.

Chương 3

Phương pháp tiếp cận

3.1 Xây dựng tập dữ liệu

3.1.1 Xác định yêu cầu bài toán

Bài toán đặt ra là nhận diện người trong khung hình có đang đeo các thiết bị bảo hộ cá nhân (tiếng Anh: personal protective equipment) hay không. Các thiết bị bảo hộ cá nhân được chọn để nhận diện là: mũ bảo hộ, áo bảo hộ và khẩu trang.



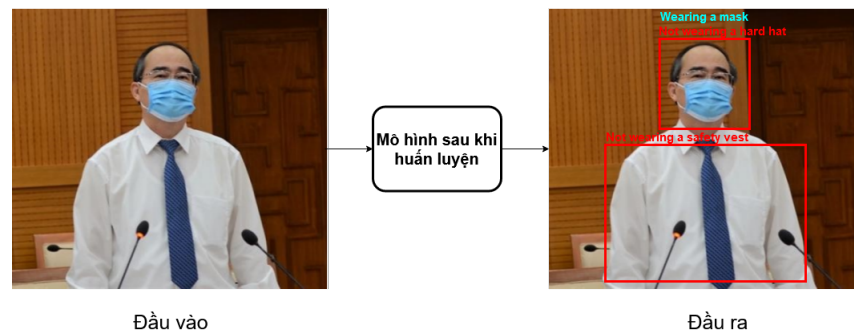
Hình 3.1: (1) Mũ bảo hộ, (2) Áo bảo hộ, (3) Khẩu trang

Mục tiêu đầu ra của hệ thống là có thể xác định được vị trí đầu người và thân người và phân loại việc sử dụng các thiết bị bảo hộ cá nhân đối với các vật thể đã được phát hiện. Ứng với mỗi thiết bị, vật thể sẽ được phân loại thành hai trạng thái, một là *Wearing* - *Mặc*, hai là *Not wearing* - *Không mặc*.

- Wearing a hardhat

- Not wearing a hardhat
- Wearing a safety vest
- Not wearing a safety vest
- Wearing a mask
- Not wearing a mask

Hình 3.2 minh họa đầu vào đầu ra mong muốn của hệ thống.



Hình 3.2: Kết quả nhận dạng mong muốn.

Do vậy tập dữ liệu sẽ được xây dựng với các hình ảnh chứa con người và các nhãn được đánh đúng với mong muốn của đầu ra.

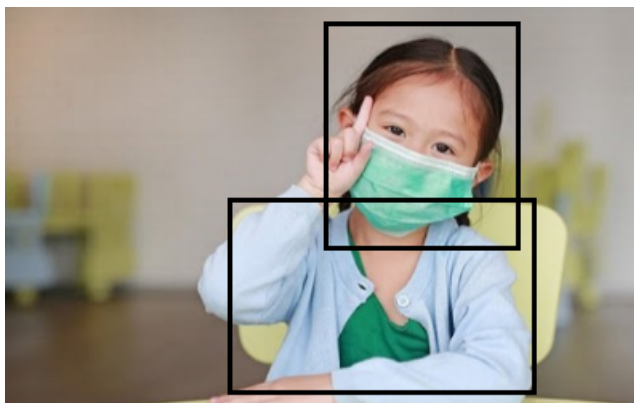
3.1.2 Thu thập hình ảnh và dán nhãn

Tập dữ liệu gồm 11586 hình trong đó

- 3541 hình được lấy từ tập dữ liệu *Hardhat and Safety Vest Image for Object Detection*
- 3174 hình được lấy từ tập dữ liệu *GDUT-HWD*
- 4871 hình được lấy từ công cụ tìm kiếm hình ảnh *Google image*

Các hình được dán nhãn bằng phần mềm *LabelImg*. Mỗi hình sẽ có tương ứng một tệp tin văn bản với đuôi *.txt*. Bên trong tệp tin này là các nhãn được đánh dấu bằng định dạng của YOLO với các thông tin gồm *object class id* - đây là id tương ứng với thứ tự của một nhãn trong danh sách các nhãn,

x và y là tọa độ tương đối của bounding box được đánh dấu với hình, w và h là chiều rộng và chiều cao tương đối của bounding box được đánh dấu với hình, hình ??.



```
3 0.590820 0.732087 0.404297 0.448598
1 0.649414 0.323988 0.306641 0.560748
4 0.649414 0.323988 0.306641 0.560748
```

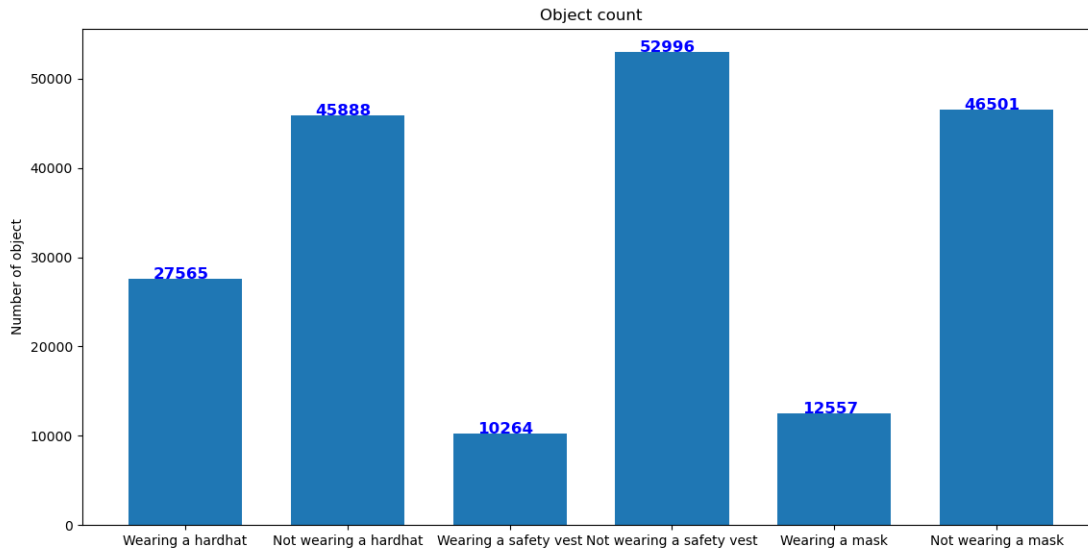
Hình 3.3: Định dạng nhãn của YOLO.

Tập dữ liệu này có tổng cộng 195771 vật thể được dán nhãn với thông kê số vật thể của từng class được thể hiện trong hình 3.4.

Việc chênh lệch lớn về số lượng các bounding box giữa các class cụ thể là các class *Not wearing* có số lượng lớn hơn rất nhiều so với các class *Wearing* tương ứng sẽ giúp bộ nhận dạng nhạy hơn với các trường hợp vi phạm trang phục bảo vệ lao động. Tuy có sự chênh lệch lớn nhưng số lượng các bounding box của mỗi class là đủ lớn để mô hình có thể học được các đặc trưng cần thiết để có thể phân loại class tốt cho một bounding box.

3.2 Huấn luyện mạng YOLOv3 sử dụng framework Darknet

Darknet là một framework được xây dựng bởi Joseph Redmon cũng là cha đẻ của YOLO, framework này được viết bằng C/C++ và được dùng để huấn luyện mô hình YOLOv3 với tập dữ liệu riêng cho từng vấn đề. Kiến trúc của Darknet đã được đề cập ở phần lý thuyết và sẽ không được nhắc lại ở chương này.



Hình 3.4: Thống kê số lượng vật thể ứng với từng class. Wearing a hardhat: 27565, Not wearing a hardhat: 45888, Wearing a safety vest: 10264, Not wearing a safety vest: 52996, Wearing a mask: 12557, Not wearing a mask: 46501.

Mô hình YOLOv3 trong luận văn này được huấn luyện trên *GoogleColab*, về bản chất môi trường trên *GoogleColab* là môi trường máy ảo chạy Linux với các thông số tại thời điểm thực hiện luận văn:

- Hệ điều hành: Ubuntu 18.04.3 LTS
- Chip xử lý: Intel 2-core Xeon 2.2GHz
- RAM: 13Gb
- HDD: 33Gb
- GPU: Tesla K80 with 12GB memory

Để có thể huấn luyện được mô hình YOLOv3 cho bộ dữ liệu riêng, ta cần thực hiện một số bước.

1. Tải framework Darknet từ github repository của AlexeyAB.

<https://github.com/AlexeyAB/darknet>

Sau đó ta chọn Clone → Download ZIP và tiến hành tải thư mục về, sau khi tải xong ta tiến hành giải nén vào một thư mục mà ta tạo sẵn.

2. Ta chép tệp tin *yolov3.cfg* trong thư mục *cfg* ra thư mục làm việc hiện tại và chỉnh sửa như sau. Đầu tiên ta sẽ sửa các giá trị *batch* là số hình trong một mini-batch mà ta muốn dùng để huấn luyện, *subdivision* là thông số để chia nhỏ một mini-batch để đảm bảo mô hình có thể chạy trên các tài nguyên GPU khác nhau, *width* và *height* là chiều rộng và chiều cao của ảnh đầu vào, các ảnh có kích thước khác nhau sẽ được resize lại kích thước này trước khi được đưa vào để huấn luyện. *max_batches* là tổng số batch mà mô hình sẽ chạy qua, đây còn được gọi là số *iteration*. *steps* sẽ có giá trị lần lượt bằng 80% và 90% của *max_batches*.

```
# [net]
# Testing
# batch=1
# subdivisions=1
# Training
batch=64
subdivisions=64
width=608
height=608
channels=3
momentum=0.9
decay=0.0005
angle=0
saturation = 1.5
exposure = 1.5
hue=.1

learning_rate=0.001
burn_in=1000
max_batches = 22000
policy=steps
steps=17600,19800
scales=.1,.1
```

Sau đó tại các dòng 610, 696 và 783 ta sẽ thay số *classes* bằng sáu, chính là số class của bài toán. Đồng thời ta sẽ sửa số *filters* của lớp convolution ngay phía trên theo công thức $3 \times (5 + C)$ với C là số class, khi $C = 6$ ta có *filters* = 33.


```

[convolutional]
size=1
stride=1
pad=1
filters=33
activation=linear

[yolo]
mask = 0, 1, 2
anchors = 10, 13, 16, 30, 33, 23, 30, 61, 62, 45, 59, 119, 116, 90, 156, 198, 373, 326
classes=6
num=9
jitter=.3
ignore_thresh = .7
truth_thresh = 1
random=1

```

3. Sửa các thông số ở đầu của tệp *Makefile* như sau.

```

GPU=1
CUDNN=0
CUDNN_HALF=0
OPENCV=1
AVX=0
OPENMP=0
LIBSO=0
ZED_CAMERA=0 # ZED SDK 3.0 and above
ZED_CAMERA_v2_8=0 # ZED SDK 2.X

```

4. Tạo tệp có tên *obj.names* chứa tên các class.

```

Wearing a hardhat
Not wearing a hardhat
Wearing a safety vest
Not wearing a safety vest
Wearing a mask
Not wearing a mask

```

5. Tạo hai tệp text, *train.txt* và *val.txt* chứa đường dẫn đến các hình ảnh. *train.txt* sẽ được dùng để huấn luyện còn *val.txt* sẽ được dùng để validate mô hình trong quá trình huấn luyện. Việc chia tập dữ liệu được thực hiện ngẫu nhiên. Có 900 hình được dùng để validate và 10686 hình được dùng để huấn luyện. Việc chia này được thực hiện bằng một chương trình Python.

```
import os
import glob
import cv2
import random

basenames = [os.path.basename(x) for x in glob.glob("./data/objects/*.jpg")]
basenamesNotEmpty = []

for name in basenames:
    if "empty" not in name:
        basenamesNotEmpty.append(name)

train = random.sample(basenames, len(basenames))
valid = random.sample(basenamesNotEmpty, 900)

with open("./val.txt", "w") as f:
    for name in valid:
        f.write("data/objects/"+name+"\n")

with open("./train.txt", "w") as f:
    for name in train:
        if name not in valid:
            f.write("data/objects/"+name+"\n")
```

6. Tạo tệp có tên *obj.data* chứa tên các class.

```
classes = 6  
train = train.txt  
valid = val.txt  
names = obj.names  
backup = backup/
```

Chương 4

Phân tích kết quả

In this chapter, ...

Chương 5

Kết luận

In this chapter, ...