ĐỀ TÀI NHẬN DIỆN MANG GIÀY HAY KHÔNG MANG GIÀY VÀ LOẠI GIÀY (3 LOẠI GIÀY)

Phạm Trần Anh Tuấn – 19146298
Email: 19146298@student.hcmute.edu.vn
Môn học: Trí tuệ tạo
Nhóm: 03CLC
Mã học phần: 212ARIN337629

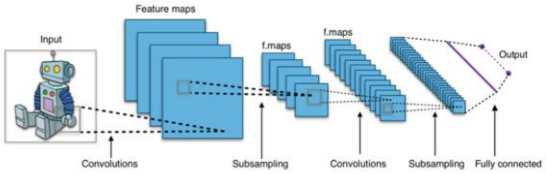
Trí tuệ nhân tạo ngày nay không còn là một thuật ngữ mới mẻ hay xa lạ nữa, mà ngày nay việc máy móc hiện hữu bên cạnh cũng như gần gũi hơn với con người. Trí tuệ nhân tạo có thể biến những thứ không thể thành có thể, biến thứ không tưởng thành hiện thực. Nó phát triển đáng kinh ngạc và dần dần thay thế con người những công việc nặng nhọc cũng như các công việc khó khăn. Những căn nhà thông minh, những robot chẩn đoán bệnh hay hoạt động trong lĩnh vực quân sự, những chiếc xe không người lái,.. Tương lai sẽ còn hơn thế nữa nó có thể phát triển đến chóng mặt hơn cả những gì ta nghĩ. Đề tài này sử dụng CNN để nhận diện các hang giày và không mang giày. Nhằm mục đích sử dụng trong các buổi hội nghị hoặc là khi trong môi trường học vấn, phát hiện rằng các khách tham dự cũng như là các học sinh sinh viên có đáp ứng được yêu cầu đồng phục giày đầy đủ hay không. CNN là tên viết tắt của từ Convolutional Neural Network (hay còn gọi là CNNs_mang nơ ron tích chập). Đây là một trong những mô hình Deep Learning vô cùng tiên tiến. CNN sẽ cho phép bạn xây dựng các hệ thống thông minh với độ chính xác vô cùng cao. Hiện nay, CNN được ứng dụng rất nhiều trong những bài toán nhận dạng object trong ảnh. CNN phân loại hình ảnh bằng cách lấy 1 hình ảnh đầu vào, xử lý và phân loại nó theo các hạng mục nhất định (Ví dụ: Chó, Mèo, Hổ, ...). Máy tính coi hình ảnh đầu vào là 1 mảng pixel và nó phụ thuộc vào độ phân giải của hình ảnh. Dựa trên độ phân giải hình ảnh, máy tính sẽ thấy H x W x D (H: Chiều cao, W: Chiều rộng, D: Độ dày). Ví dụ: Hình ảnh là mảng ma trận RGB 6x6x3 (3 ở đây là giá trị RGB). Về kỹ thuật, mô hình CNN để training và kiểm tra, mỗi hình ảnh đầu vào sẽ chuyển nó qua 1 loạt các lớp tích chập với các bộ lọc (Kernals), tổng hợp lại các lớp được kết nối đầy đủ (Full Connected) và áp dụng hàm Softmax để phân loại đối tượng có giá trị xác suất giữa 0 và 1. Hình dưới đây là toàn bộ luồng CNN để xử lý hình ảnh đầu vào và phân loại các đối tượng dựa trên giá trị.

1. Tổng quan

Hiện nay với công nghệ ngày càng phát triển cuộc sống con người ngày càng được cải thiện, điều đó dẫn tới việc các vật dụng xung quanh chúng ta ngày càng được nâng cấp lên, giày cũng là một trong số đó. Giày đã phát triển từ lâu đời và dần dần chứng minh được sự hữu ích trong cuộc sống và thời trang. Ngày nay, khi đi học cũng như tham gia các cuộc hội nghị thì hầu như là các đại biểu cũng như là học sinh sinh viên thường mang giày để tham dự và đi học, việc này góp phần làm cho chính bản thân mỗi người thêm đẹp khi đi và cũng góp phần cho buổi hội nghị là cảnh quan trường học, lớp học thêm phần chỉn chu và lịch sự hơn. Thế nhưng vẫn có một vài thành phần là vẫn mang dép lê, dép quai kẹp,… với lí do bao biện là cho thoải mái. Thế nhưng nó làm cho tổng quan không có thẩm mỹ và kém lịch sự đi một chút. Chính vì do đó nhằm để quan sát cũng như có thể sử dụng trong việc check in đầu vào khi tham dự cũng như khi đi học vào lớp rằng xem bạn có đáp ứng đủ yêu cầu giày dép trang phục khi đi hay không thì đề tài này có tác dụng nhận diện chính xác và giúp phần đỡ hơn cho các người check in khi kiểm tra đồng phục giày dép tham dự.

Mạng CNN là một tập hợp các lớp Convolution chồng lên nhau và sử dụng các hàm nonlinear activation như ReLU và tanh để kích hoạt các trọng số trong các node. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Mỗi một lớp sau khi thông qua các hàm kích hoạt sẽ tạo ra các thông tin trừu tượng hơn cho các lớp tiếp theo. Trong mô hình mạng truyền ngược (feedforward neural network) thì mỗi neural đầu vào (input node) cho mỗi neural đầu ra trong các lớp tiếp theo. Mô hình này gọi là mạng kết nối đầy đủ (fully connected layer) hay mạng toàn vẹn (affine layer). Còn trong mô hình CNNs thì ngược lại. Các layer liên kết được với nhau thông qua cơ chế convolution. Layer tiếp theo là kết quả convolution từ layer trước đó, nhờ vậy mà ta có được các kết nối cục bộ. Như vậy mỗi neuron ở lớp kế tiếp sinh ra từ kết quả của filter áp đặt lên một vùng ảnh cục bộ của neuron trước đó. Mỗi một lớp được sử dụng các filter khác nhau thông thường có hàng trăm hàng nghìn filter như vậy và kết hợp kết quả của chúng lại. Ngoài ra có một số layer khác như pooling/subsampling layer dùng để chắt lọc lại các thông tin hữu ích

hơn (loại bỏ các thông tin nhiễu). Trong quá trình huấn luyện mạng (traning) CNN tự động học các giá trị qua các lớp filter dựa vào cách thức mà bạn thực hiện. Ví dụ trong tác vụ phân lớp ảnh, CNNs sẽ cố gắng tìm ra thông số tối ưu cho các filter tương ứng theo thứ tự raw pixel > edges > shapes > facial > high-level features. Layer cuối cùng được dùng để phân lớp ảnh.



Hình 1.1. Ảnh minh họa các lớp CNN.

2. Tài liệu và giải pháp

2.1 Thu thập dữ liệu, khó khăn và thách thức

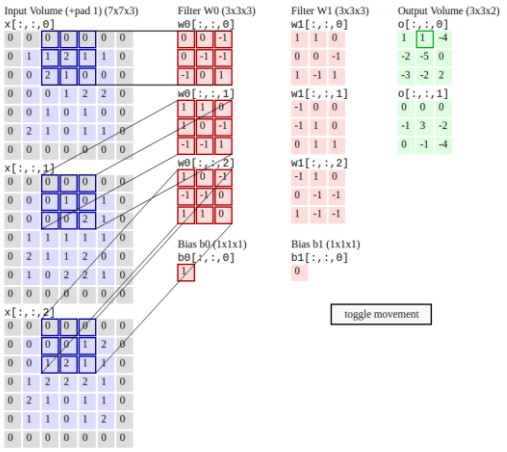
Dựa trên một phần dữ liệu có sẵn trên Kaggle và chủ yếu là các tấm ảnh chỉnh sửa lại trên các web bán hang của các hãng cụ thể là Nike, Adidas và Converse. Các ảnh trong tập xử lí dữ liệu độ nhiễu khá cao vì chưa qua tiền xư lí dữ liệu. Và một phần là do các hãng giày có kiểu giày khá là giống nhau nên khó có thể phân biệt được các loại.

2.2 Datasets và tiền xử lí dữ liệu

Đầu tiên ta cắt ảnh từ trang web bán hàng của các hãng. Sau đó ta gắn nhãn cho các sản phẩm cụ thể là giày lần lượt là 0, 1, 2, 3. Sau đó chỉnh lại kích thước các ảnh đầu vào là 150x150. Với 3 kênh màu là đỏ xanh lá và xang dương.

Như trình bày ở trên, Convolutional Neural Network là một trong những phương pháp chính khi sử dụng dữ liệu về ảnh. Kiến trúc mạng này xuất hiện do các phương pháp xử lý dữ liệu ảnh thường sử dụng giá trị của từng pixel. Vậy nên với một ảnh có giá trị kích thước 100x100 sử dụng kênh RGB ta có tổng cộng ta có 100 * 100 * 3 bằng 30000 nút ở lớp đầu vào. Điều đó kéo theo việc có một số lượng lớn weight và bias dẫn đến mạng nơ-ron trở nên quá đồ sộ, gây khó khăn cho việc tính toán. Hơn nữa, chúng ta có thể thấy rằng thông tin của các pixel thường chỉ chịu tác động bởi các pixel ngay gần nó, vậy nên việc bỏ qua một số nút ở tầng đầu vào trong mỗi lần huấn luyện sẽ không làm giảm độ chính xác của mô hình. Vậy nên người ta sử dụng cửa số tích chập nhằm giải quyết vấn đề số lượng tham số lớn mà vẫn trích xuất được đặc trưng của ảnh.

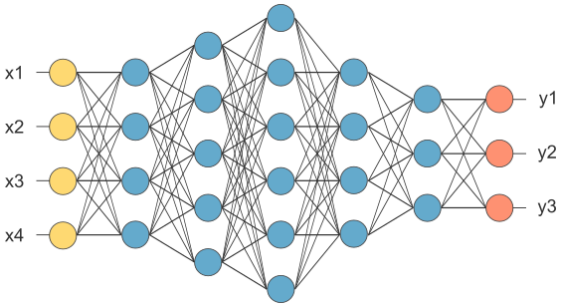
Lớp tích chập đầu tiên Convolution Layer là lớp đầu tiên trích xuất các đặc tính từ hình ảnh. Tham số lớp này bao gồm một tập hợp các bộ lọc có thể học được. Các bộ lọc đều nhỏ thường có kích cỡ hai chiều đầu tiên khoảng 3x3 hoặc 5x5, ... và có độ sâu bằng với độ sâu của đầu vào đầu vào. Bằng cách trượt dần bộ lọc theo chiều ngang và dọc trên ảnh, chúng thu được một Feature Map chứa các đặc trưng được trích xuất từ trên hình ảnh đầu vào.



Hình 2.1. Ảnh minh họa.

Lớp tiếp theo là 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. Trong quá trình này, quy tắc về stride và padding áp dụng như phép tính convolution trên ảnh.

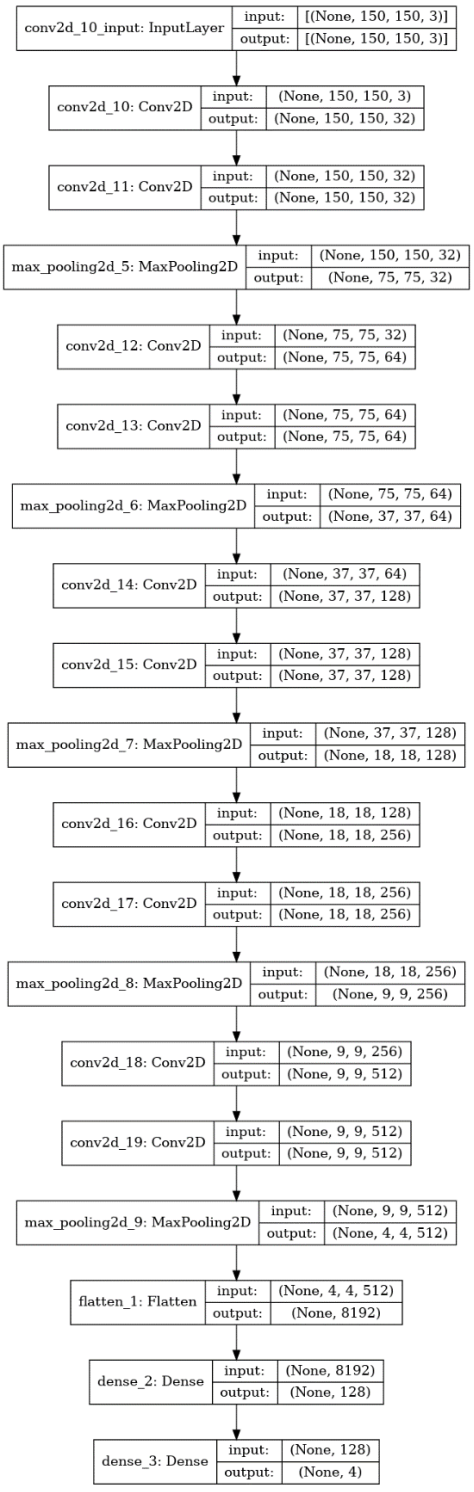
Sau khi ảnh được truyền qua nhiều convolutional layer và pooling layer thì model đã học được tương đối các đặc điểm của ảnh thì tensor của output của layer cuối cùng sẽ được là phẳng thành vector và đưa vào một lớp được kết nối như một mạng nơ-ron. Với FC layer được kết hợp với các tính năng lại với nhau để tạo ra một mô hình. Cuối cùng sử dụng softmax hoặc sigmoid để phân loại đầu ra.



Hình 2.2. Ảnh minh họa.

2.3 Huấn luyện mô hình

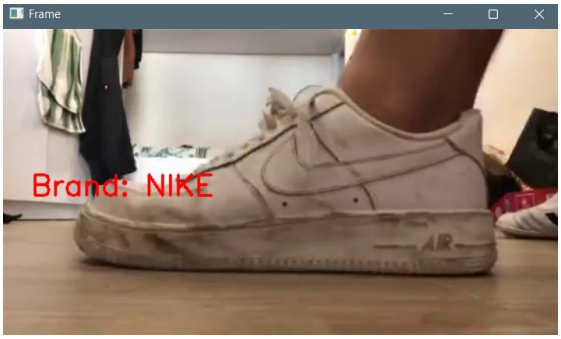
Sau khi hoàn thành các bước tiền xử lí dữ liệu ta bắt đầu huấn luyện cho máy học. Có tất cả các lớp sau



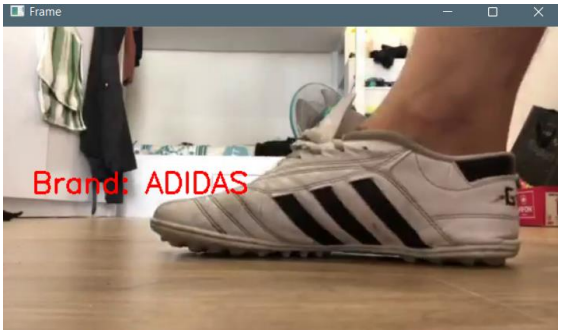
Hình 2.3. Ảnh các lớp.

3. Kết quả

Hình 3.1. Ảnh test thử giày Adidas.



Hình 3.2. Ảnh khi chạy thời gian thực loại Nike.

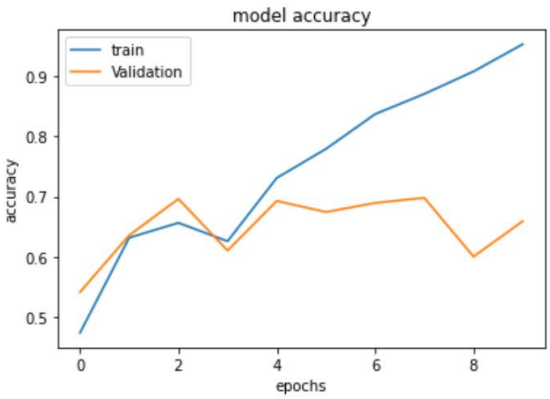


Hình 3.3. Ảnh chạy thời gian thực loại Adidas.

Sau khi thử nghiệm trên các ảnh, video và thời gian thực thì ta nhận thấy rằng giày Adidas và Nike nhận khá tốt và No_shoe cũng ổn chỉ có là Giày hãng Converse là không được chính xác lắm. Hình ảnh sau đây là các lớp trong model.

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_30 (Conv2D)           (None, 150, 150, 32)      896

conv2d_31 (Conv2D)           (None, 150, 150, 32)      9248

max_pooling2d_15 (MaxPoolin  (None, 75, 75, 32)        0
g2D)

conv2d_32 (Conv2D)           (None, 75, 75, 64)        18496

conv2d_33 (Conv2D)           (None, 75, 75, 64)        36928

max_pooling2d_16 (MaxPoolin  (None, 37, 37, 64)        0
g2D)

conv2d_34 (Conv2D)           (None, 37, 37, 128)       73856

conv2d_35 (Conv2D)           (None, 37, 37, 128)       147584

max_pooling2d_17 (MaxPoolin  (None, 18, 18, 128)       0
g2D)

conv2d_36 (Conv2D)           (None, 18, 18, 256)       295168

conv2d_37 (Conv2D)           (None, 18, 18, 256)       590080

max_pooling2d_18 (MaxPoolin  (None, 9, 9, 256)         0
g2D)

conv2d_38 (Conv2D)           (None, 9, 9, 512)         1180160

conv2d_39 (Conv2D)           (None, 9, 9, 512)         2359808

max_pooling2d_19 (MaxPoolin  (None, 4, 4, 512)         0
g2D)

flatten_3 (Flatten)          (None, 8192)              0

dense_6 (Dense)              (None, 128)               1048704

dense_7 (Dense)              (None, 4)                 516

=================================================================
Total params: 5,761,444
Trainable params: 5,761,444
Non-trainable params: 0
```

Hình 3.4. Ảnh các lớp.



Hình 3.5. Ảnh quá trình học.

4. Kết luận và các hạn chế

Tổng quan lại mô hình học chưa được ổn định cho lắm. Nhưng quan trọng là mô hình nhận diện thời gian thực khá là tốt, nhận diện được mang giày hay không mang giày và không mang giày và biết được các loại giày.

Nhưng bên cạnh đó các hạn chế của đề tài như:

- Hạn chế thứ nhất: Giới hạn về không được sử dụng các model train sẵn như Yolo, VGG,… điều này làm cho việc nhận diện thời gian thực giảm độ chính xác xuống. Thay vì sẽ có khung xanh để bắt đôi giày của chúng ta thì phải để camera dưới đất và khá sát giày thì độ chính xác mới cao hơn.
    - Hạn chế thứ hai: Tập dữ liệu train chưa thực sự hoàn hảo, cụ thể như ở trên biểu đồ học của máy ta thấy rằng lúc đầu máy bám khá sát đường train nhưng càng về sau thì không còn được bám sát như lúc đầu nữa(overfiting). Việc chưa có kinh nghiệm soạn tập dữ liệu cũng là một hạn chế.
    - Hạn chế thứ ba: Chưa có tối ưu hóa hết được thuật toán trong nhận diện thời gian thực khiến cho việc nhận diện đôi lúc không chính xác.

Hướng phát triển Có thể sử dụng them các model được huấn luyện sẵn để tối

ưu hóa được nhận diện thời gian thực. Bên cạnh đó cải thiện thêm phần dữ liệu huấn luyện được tối ưu đầy đủ và chi tiết hơn các đặc trưng của giày để máy có thể học hiệu quả và được tối ưu hơn.

Chung quy lại để hoàn thành tốt dự án này chúng em đã được sự hướng dẫn của thầy PGS. TS Nguyễn Trường Thịnh và các anh chị trợ giảng đồng hành suốt kỳ và thời gian thực hiện dự án này. Qua đó chúng em rút ra được nhiều kinh nghiệm cũng như thêm được nhiều trải nghiệm thú vị và bổ ích.

Tài liệu tham khảo

- https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/
- Các file PowerPoint từ PGS.TS Nguyễn Trường Thịnh