- **Nguyễn Văn Long 1712024**
- **Nguyễn Thị Bích Lan 1711884**

Đề tài: Nhận diện khuôn mặt

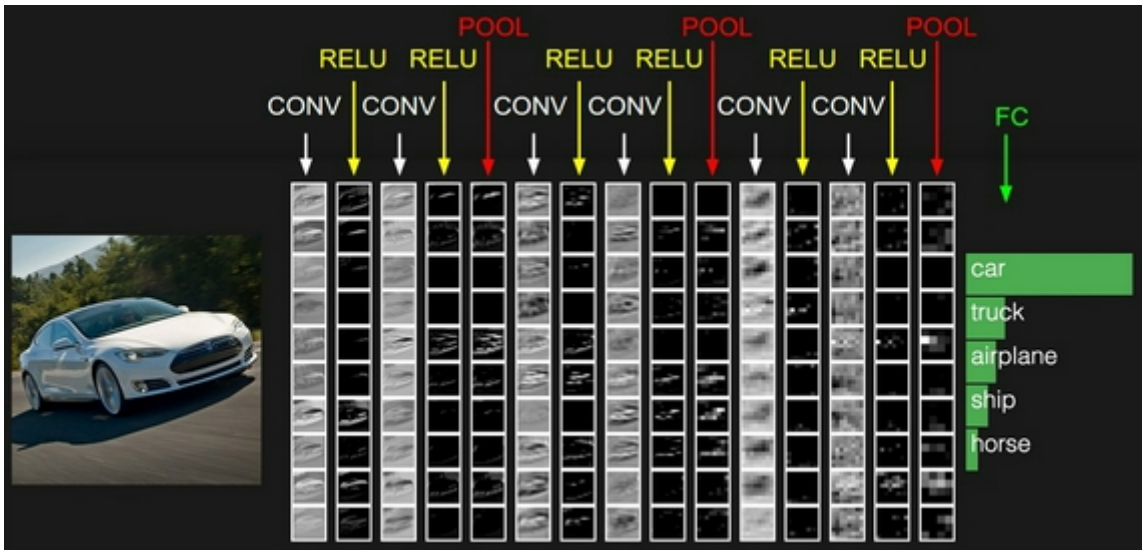
GVHD: Trần Quốc Tiến Dũng

Convolutional neural network (Mạng nơ-ron tích chập)

1.Máy tính làm gì với CNN?

CNN cho phép các máy tính có khả năng “nhìn” và “phân tích”, nói 1 cách dễ hiểu, CNN được sử dụng để nhận dạng hình ảnh bằng cách đưa hình ảnh đó qua nhiều layer với một bộ lọc tích chập để sau cùng có được một điểm số nhận dạng đối tượng.

CNN có 02 phần chính: Lớp trích xuất đặc trưng của ảnh (Conv, Relu và Pool) và Lớp phân loại (FC và softmax).

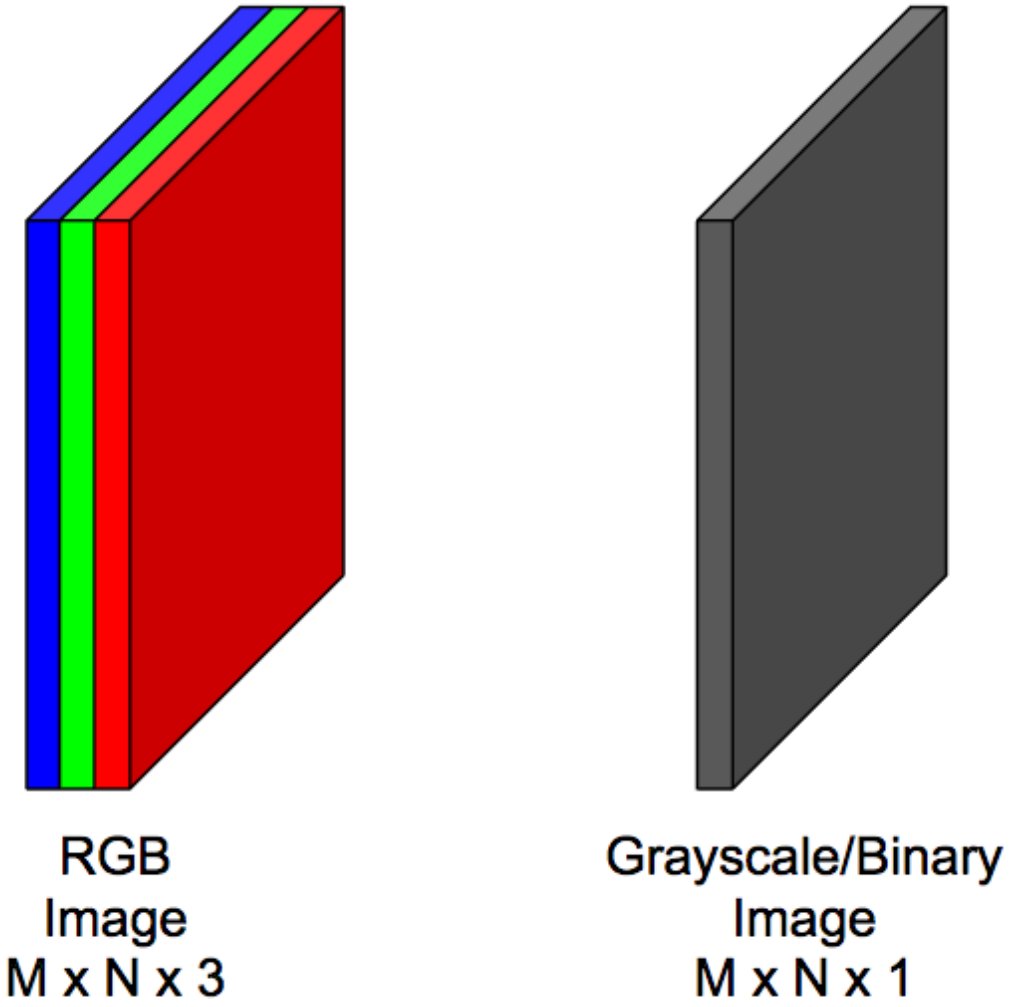


Đầu vào (dữ liệu training):

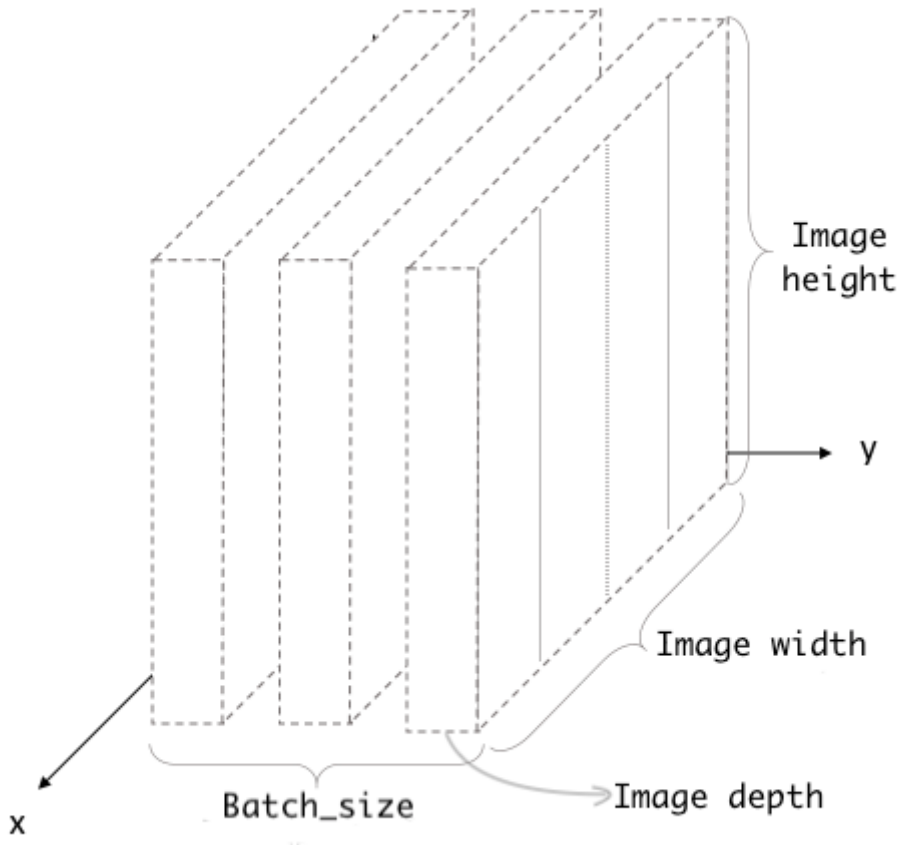
Input đầu vào là một bức ảnh được biểu diễn bởi ma trận pixel với kích thước: [W x H x D]

- W: chiều rộng
- H: chiều cao
- D: Là độ sâu, hay dễ hiểu là số lớp màu của ảnh.

Ví dụ ảnh RGB sẽ là 3 lớp ảnh Đỏ, Xanh Dương, Xanh.



Kích thước input và output của môn hình CNN:



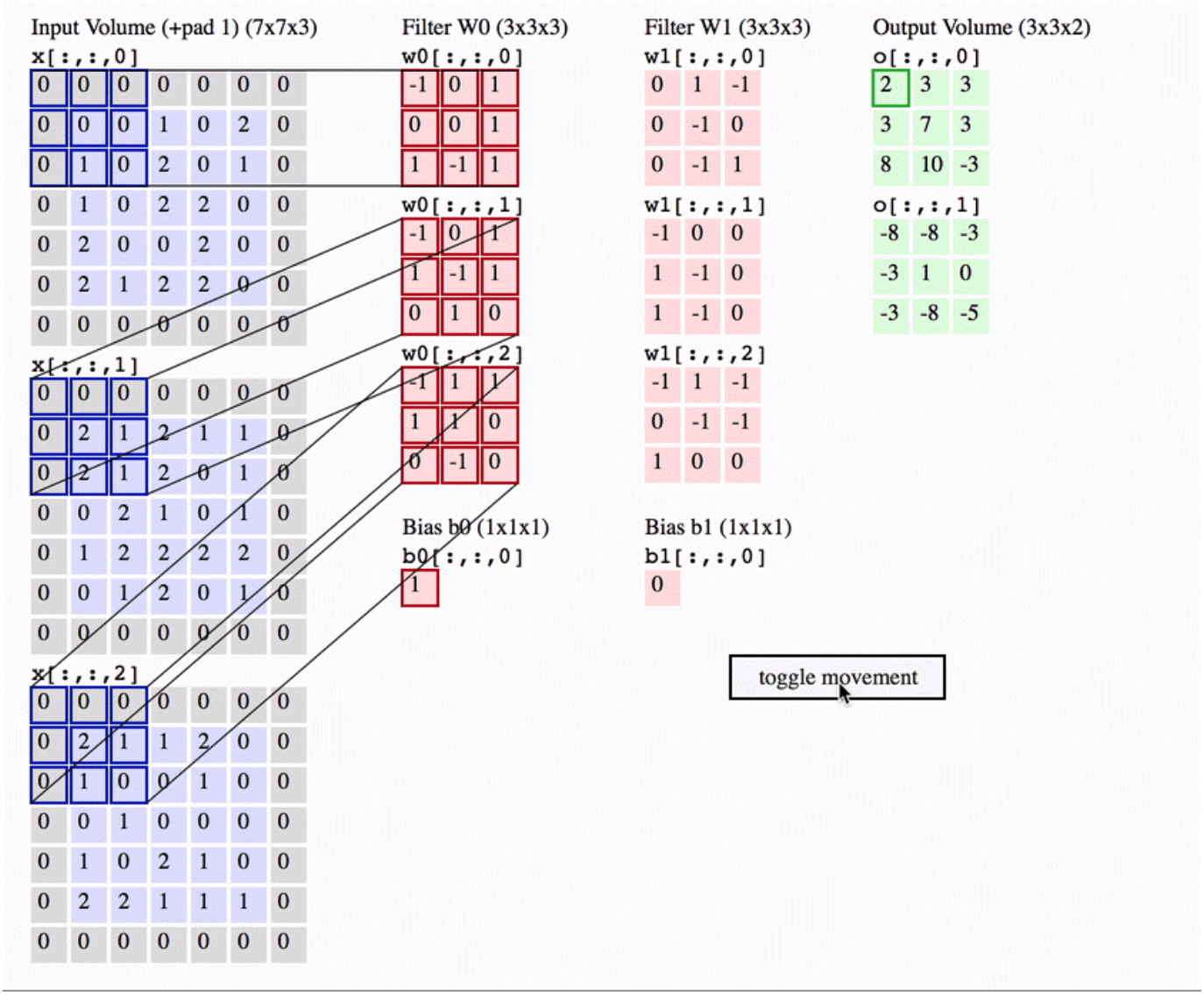
Batch_size là số tấm ảnh chúng ta đưa vào để train. Image_depth là số ma trận biểu thị cường độ sáng các màu. Ví dụ:

- Đối với n tấm ảnh RBG thì batch_size là n, Image_depth là 3
- Đối với n tấm ảnh Gray thì Batch_size tương tự,nhưng Image_depth là 1.

2. Conv Layer:

Mục tiêu của các lớp tích chập là trích xuất các đặc trưng của ảnh đầu vào.

Ví dụ:



Tại sao CNNs cho ra mô hình với độ chính xác rất cao?

CNNs có tính bất biến và tính kết hợp cục bộ (Location Invariance and Compositionality). Với cùng một đối tượng, nếu đối tượng này được chiếu theo các góc độ khác nhau (translation, rotation, scaling) thì độ chính xác của thuật toán sẽ bị ảnh hưởng đáng kể. Pooling layer sẽ cho bạn tính bất biến đối với phép dịch chuyển (translation), phép quay (rotation) và phép co giãn (scaling).

Tính kết hợp cục bộ cho ta các cấp độ biểu diễn thông tin từ mức độ thấp đến mức độ cao và trừu tượng hơn thông qua convolution từ các filter. Đó là lý do tại sao CNNs cho ra mô hình với độ chính xác rất cao. Cũng giống như cách con người nhận biết các vật thể trong tự nhiên. Ta phân biệt được một con chó với một con mèo nhờ vào các đặc trong từ mức độ thấp (có 4 chân, có đuôi) đến mức độ cao (dáng đi, hình thể, màu lông)

Kích thước của các bộ lọc và ý nghĩa của chúng:

Ảnh đầu vào được cho qua một bộ lọc chạy dọc bức ảnh. Bộ lọc có kích thước là (3x3 hoặc 5x5) và áp dụng phép tích vô hướng để tính toán, cho ra một giá trị duy nhất. Đầu ra của phép tích

chập là một tập các giá trị ảnh được gọi là mạng đặc trưng (features map).

Thực chất, ở các layer đầu tiên, phép tích chập đơn giản là phép tìm biên ảnh. Còn nếu không thì bạn chỉ cần hiểu sau khi cho qua bộ lọc nó sẽ làm hiện lên các đặc trưng của đối tượng trong ảnh như đường vẽ xung quanh đối tượng, các góc cạnh,v.v.., và các layer tiếp theo sẽ lại trích xuất tiếp các đặc trưng của đặc trưng của các đối tượng đó, việc có nhiều layer như vậy cho phép chúng ta chia nhỏ đặc trưng của ảnh tới mức nhỏ nhất có thể. Vì thế mới gọi là mạng đặc trưng.

- Filter, Kernel hay Feature Detector đều là cách gọi của ma trận lọc (như mình đã đề cập ở trên). Thông thường, ở các lớp đầu tiên của Conv Layer sẽ có kích thước là [5x5x3] hoặc là [3x3x3] tùy theo mục đích sử dụng của từng người.
- Ưu -nhược:

Search this file...

	Smaller Filter Sizes	Larger Filter Sizes
2	Two 3x3 kernels result in an image size reduction by 4	one 5x5 kernel results in same reduction.
3	We have used (3x3 + 3x3) = 18 weights.	We used (5x5) = 25 weights.
4	So, we get lower no. of weights but more layers.	Higher number of weights but lesser layers.
5	Therefore, computationally efficient.	And, this is computationally expensive.
6	With more layers, it learns complex, more non-linear features.	With less layers, it learns simpler non linear feature
7	With more layers, it necessitates the need for larger memory.	And, it will use less memory for backpropogation.

small_large_filter_size_example_compare.csv hosted with ❤ by GitHub view raw

- Mỗi 1 kernel lọc ra 1 loại đặc trưng.
- Convolved Feature, Activation Map hay Feature Map là đầu ra của ảnh khi cho bộ lọc chạy hết bức ảnh với phép tích vô hướng.
- Receptive field là vùng ảnh được chọn để tính tích chập, hay bằng đúng cái kích thước của bộ lọc.
- Depth là số lượng bộ lọc. Lưu ý: ở đây là số lượng bộ lọc (filter) chứ không phải số lượng kênh màu RBG như ở trên.
- Stride được hiểu là khoảng cách dịch chuyển của bộ lọc sau mỗi lần tính. Ví dụ khi stride=2. Tức sau khi tính xong tại 1 vùng ảnh, nó sẽ dịch sang phải 2 pixel. Tương tự với việc dịch xuống dưới.
- Zero-Padding là việc thêm các giá trị 0 ở xung quanh biên ảnh, để đảm bảo phép tích chập được thực hiện đủ trên toàn ảnh.

Vậy kích thước đầu ra của ảnh với mỗi layer được tính như thế nào?

Giả sử ảnh đầu ra là [W2 x H2 x D2]

Thì: $H2 * W2 * D2 = (\frac{H-F+2P}{S} + 1) * (\frac{W-F+2P}{S} + 1) * K$

Trong đó:

- [W1xH1xD1]: Kích thước đầu vào
- F: Kích thước bộ lọc Kernel
- S: giá trị Stride

- P: số khung zero-padding thêm vào viền ảnh
- K: Số lượng bộ lọc (Depth)

3. ReLU Layer:

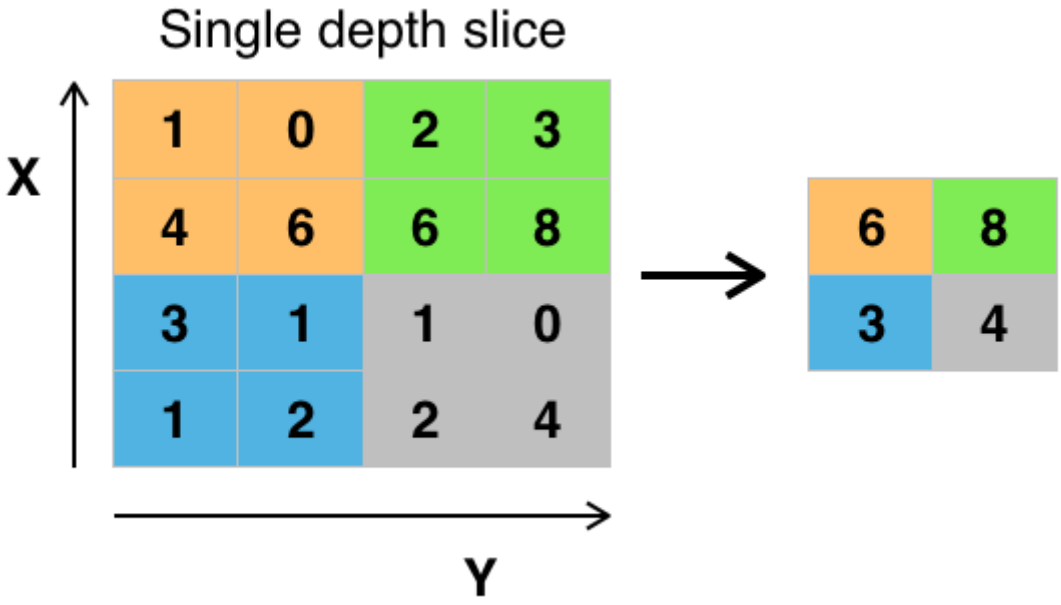
ReLU layer áp dụng hàm kích hoạt (activation function) max(0,x) lên đầu ra của Conv Layer, có tác dụng đưa các giá trị âm về thành 0. Layer này không thay đổi kích thước của ảnh và không có thêm bất kì tham số nào.

Mục đích của lớp ReLu là đưa ảnh một mức ngưỡng, ở đây là 0. Để loại bỏ các giá trị âm không cần thiết mà có thể sẽ ảnh hưởng cho việc tính toán ở các layer sau đó.

4. Pooling Layer:

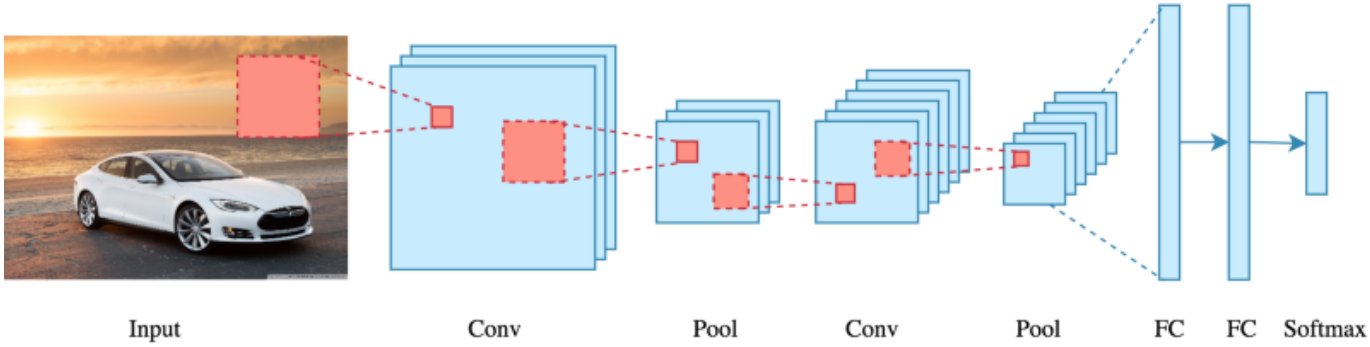
Pool Layer thực hiện chức năng làm giảm chiều không gian của đầu vào và giảm độ phức tạp tính toán của model ngoài ra Pool Layer còn giúp kiểm soát hiện tượng overffiting. Số lượng weights càng nhiều thì độ phức tạp của model càng lớn và dễ dẫn đến overfit, hoặc quá ít thì không đủ mạnh với những dữ liệu phức tạp. Vì vậy ta cần cân nhắc kỹ việc lựa chọn model sao cho phù hợp với dữ liệu huấn luyện để cho ra một model có độ chính xác cao. Thông thường, Pool layer có nhiều hình thức khác nhau phù hợp cho nhiều bài toán, ta không thể khẳng định phương pháp pooling nào là tốt nhất, vì mỗi phương pháp có điểm mạnh ở các bài toán khác nhau.

Ví dụ: Max pooling với bộ lọc 2x2 và stride = 2. Bộ lọc sẽ chạy dọc ảnh. Và với mỗi vùng ảnh được chọn, sẽ chọn ra 1 giá trị lớn nhất và giữ lại.



Thông thường max pooling có kích thước là 2 và stride=2. Nếu lấy giá trị quá lớn, thay vì giảm tính toán nó lại làm phá vỡ cấu trúc ảnh và mất mát thông tin nghiêm trọng. Vì vậy mà một số chuyên gia không thích sử dụng layer này mà thay vào đó sử dụng thêm các lớp Conv Layer và tăng số stride lên mỗi lần.

5. Fully_Connected Layer (FC):



- Sau khi kết thúc phần trích xuất đặc trưng. Ta được 1 vector đặc trưng của ảnh. Vector đó là input cho lớp fully connected layer.
- Tên tiếng viết của mạng này là Mạng liên kết đầy đủ. Tại lớp mạng này, mỗi một nơ-ron của layer này sẽ liên kết tới mọi nơ-ron của lớp trước đó.
- Đối với bài toán phân loại, đây như là 1 mạng Multi Layer Perceptron (MLP) để phân loại các tấm ảnh sau khi qua bước trích xuất đặc trưng. Tại layer cuối cùng sẽ sử dụng hàm softmax tạo ra 1 vector xác suất để phân loại các đối tượng.