

Câu 1:

- So sánh supervised và unsupervised. Ví dụ:

+ Supervised: có biến đầu vào X (input) và biến đầu ra Y (label), và sử dụng thuật toán để tìm hàm ánh xạ $Y = f(X)$.

+ Ví dụ: Nhận dạng khuôn mặt

+ Unsupervised: không biết được label Y, chỉ có dữ liệu đầu vào X. Thuật toán unsupervised sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó như clustering, dimension reduction,...

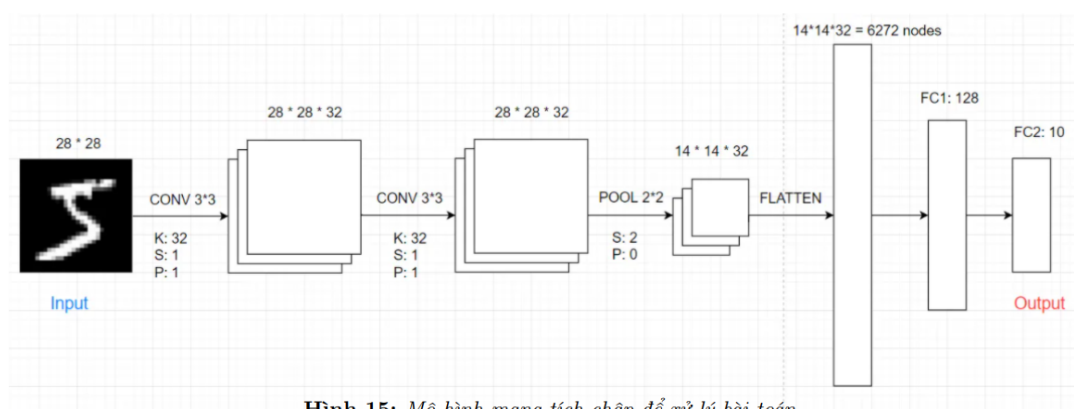
+ Ví dụ: Tìm KOL trên mạng xã hội

- Ứng dụng dùng cả supervised và unsupervised:

+ Hệ thống gợi ý sản phẩm trên sàn thương mại điện tử

+ Bài toán nhận dạng ảnh người, động vật hoặc các văn bản khoa học thu thập từ Internet thì chỉ có 1 phần là được gán nhãn, phần lớn còn lại thì không được gán nhãn. Đây là ví dụ điển hình cho sự kết hợp giữa supervised và unsupervised. Vì việc thu thập dữ liệu có nhãn tốn rất nhiều thời gian và chi phí cao. Ngoài ra có 1 số nội dung chỉ có chuyên gia mới gán nhãn được (ảnh y học).

Câu 2: Giải thích BTL, vẽ flow diagram cho BTL



Hình 15: Mô hình mạng tích chập để xử lý bài toán

- Input: MNIST là bộ cơ sở dữ liệu phổ biến về ảnh các chữ số viết tay từ 0 đến 9 bao gồm 2 tập con: tập huấn luyện gồm 60000 ảnh và tập kiểm tra gồm 10000 ảnh. Trong đó mỗi ảnh dữ liệu là một ảnh xám có kích thước 28x28

- Reshape lại dữ liệu ảnh có kích thước (N,28,28) thành kích thước (N,28,28,1) để giống kích thước mà keras yêu cầu. Xây dựng mô hình cho bài toán theo mạng tích chập như hình trên:

+ Tầng tích chập với 32 kernel kích thước 3*3 dùng hàm kích hoạt là hàm ReLU

+ Tầng tổng hợp

+ Tầng làm phẳng để chuyển kiểu dữ liệu từ tensor sang vector. Từ tensor có kích thước (14, 14, 32), số nút (node) ta có sau khi làm phẳng là $14 * 14 * 32 = 6272$ nút.

+ Tầng kết nối đầy đủ (FC) với 128 nút và dùng hàm kích hoạt là hàm sigmoid

- Output: vector softmax 10 chiều tương ứng với các giá trị 0,1,2,3,...9

Câu 3: Giả sử ta cần dự đoán kết quả xét nghiệm của 1005 bệnh nhân xem họ có bị ung thư hay không. Dưới đây là những gì mô hình của chúng ta dự đoán:

- 90 bệnh nhân bị ung thư và tất cả dự đoán này của chúng ta đều đúng.

- 915 bệnh nhân không bị ung thư nhưng thật ra có tới 910 người lại bị bệnh trong thực tế.

	Thực tế (có)	Thực tế (không)
Dự đoán (có)	90 (True Positive)	0 (False Positive)
Dự đoán (không)	910 (False Negative)	5 (True Negative)

- **True Positive (TP)**: những bệnh nhân ta **đ đoán là có** bệnh đúng là đang **mang bệnh**.
- **True Negative (TN)**: những bệnh nhân ta **đ đoán là không** có bệnh đúng là đang **khỏe mạnh**.
- **False Positive (FP)**: những bệnh nhân ta **đ đoán là có** bệnh thật ra đang **khỏe mạnh**.
- **False Negative (FN)**: những bệnh nhân ta **đ đoán là không** có bệnh thật ra đang **mang bệnh**.

Công thức tính toán:

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Công thức tổng quát:

$$\text{Sensitivity (độ nhạy)} = \frac{\text{True positive}}{\text{hàng Positive}}$$

$$\text{Specificity (độ chuyên biệt)} = \frac{\text{True negative}}{\text{hàng Negative}}$$

$$\text{Precision (độ chính xác)} = \frac{\text{True position}}{\text{cột Positive}}$$

$$\text{Negative Predictive Value} = \frac{\text{True negative}}{\text{cột Negative}}$$

$$\text{Accuracy (độ đúng đắn)} = \frac{\text{True positive} + \text{True negative}}{\text{All}}$$

$$\text{Misclassification (phân loại sai)} = \frac{\text{False positive} + \text{False negative}}{\text{All}}$$