INTRODUCTION TO DEEP LEARNING IN COMPUTER VISION

Dr. Muhammad Sajjad

R.A: Kaleem Ullah

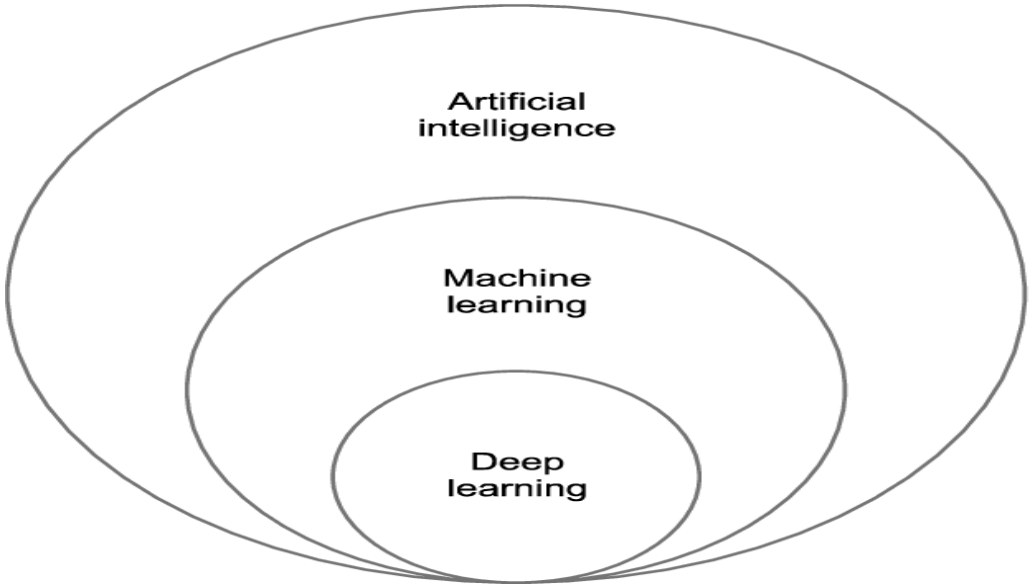
R.A: Muhammad Afaq

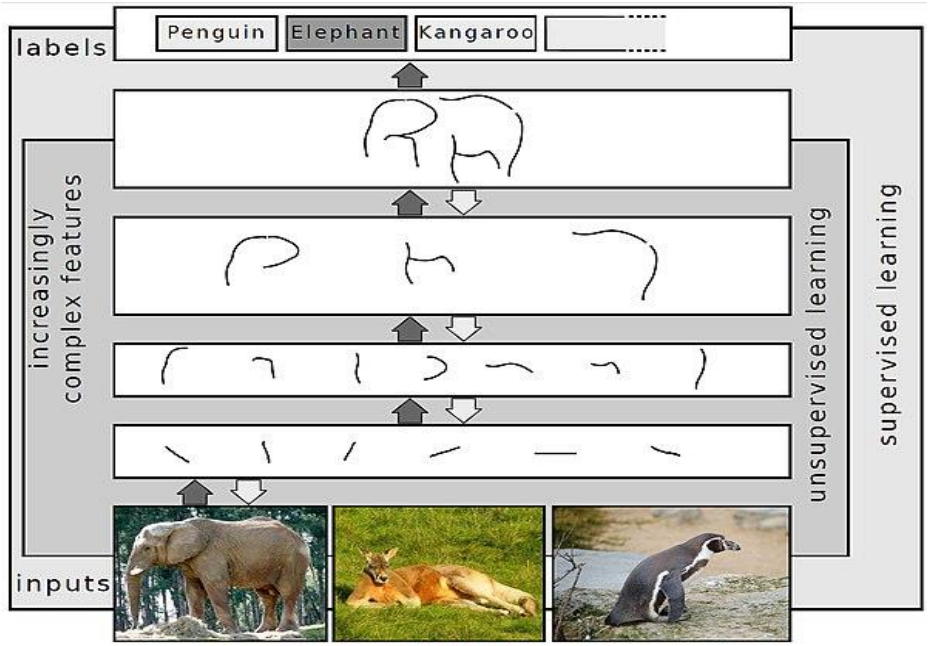
WHAT IS DEEP LEARNING

- Deep learning is a type of machine learning and Artificial Intelligence that deals with algorithms inspired by the structure and function of the human brain's neural networks.

  - The term “deep” refers to using **multiple layers** in the neural network.
  - These layers progressively extract **higher-level features** from raw input data.

  - For example:
    - In **image processing**, lower layers might detect edges.
    - Higher layers recognize more complex concepts like digits, letters, or faces.
    - Look at the image using for classification

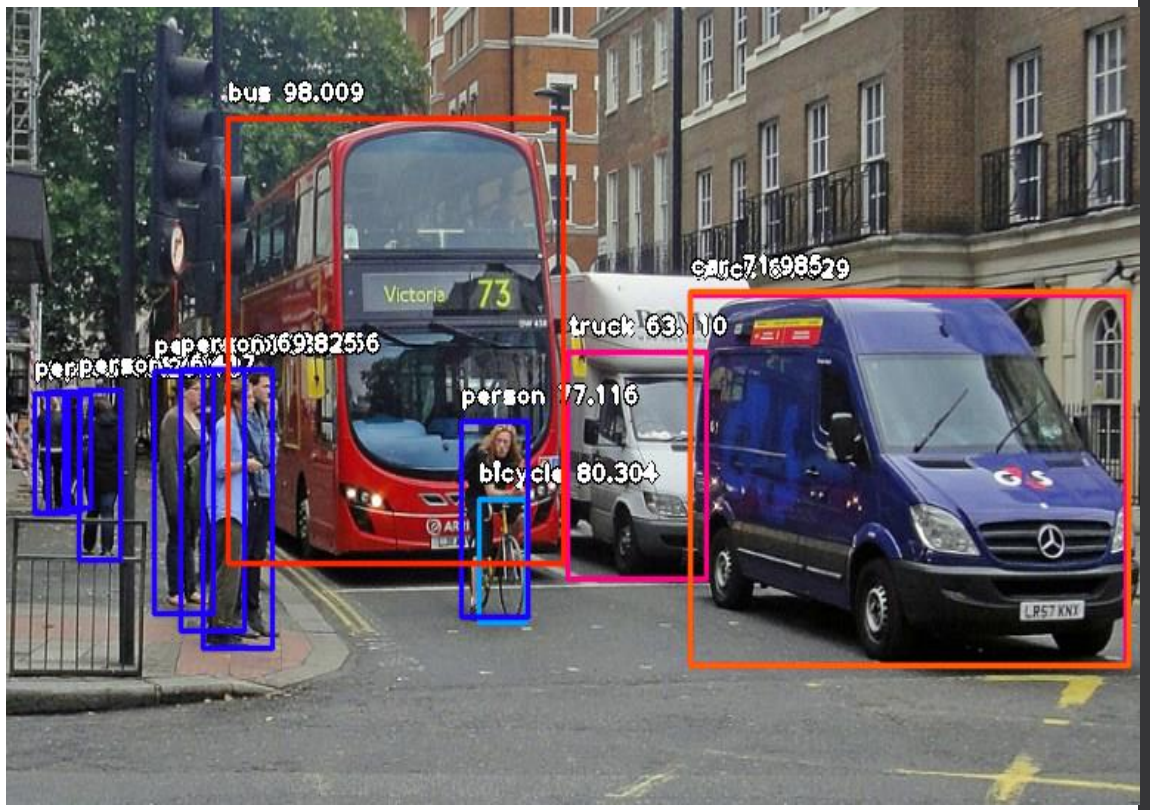




DEEP LEARNING APPLICATIONS

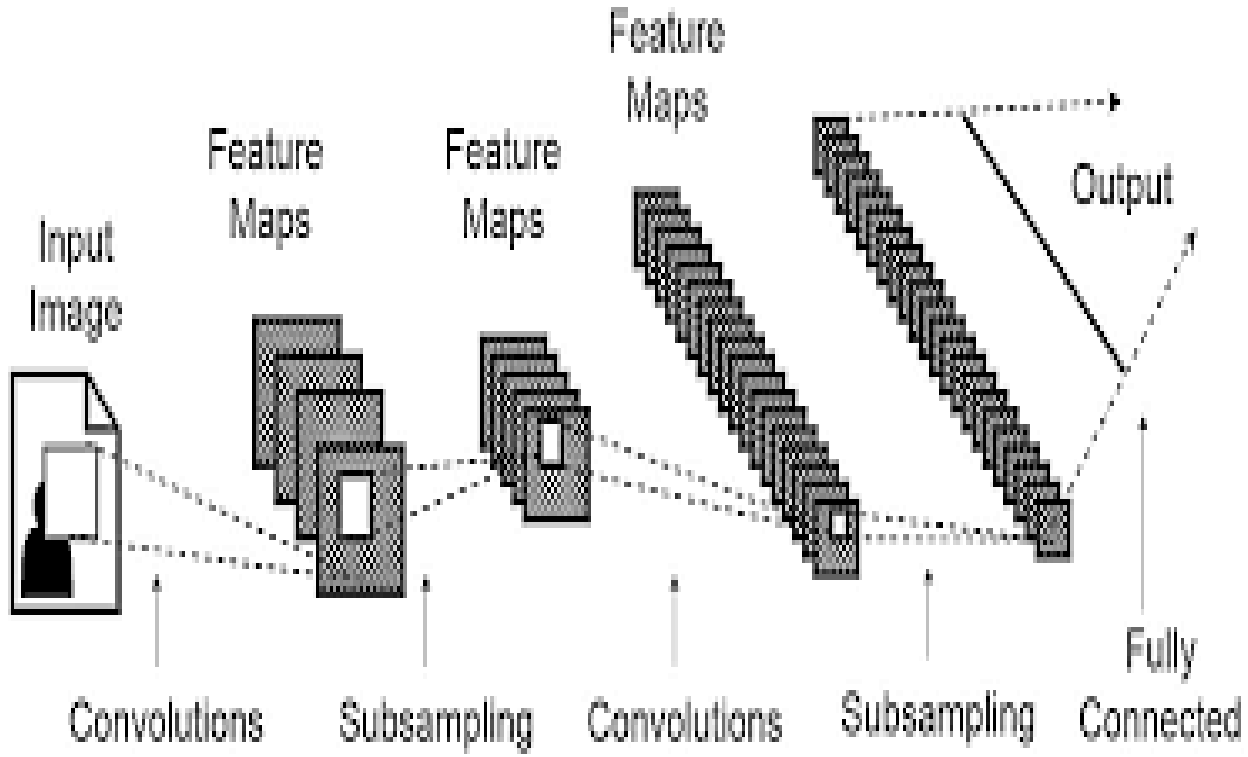
- **Applications**:

- Deep learning has transformed various fields:
  - **Computer vision**: Identifying objects, faces, and scenes in images.
  - **Medical image analysis**: Detecting diseases from medical images.
  - **Facial Analysis:** Detecting emotions, age, and gender in facial images.
  - **Architectures**: Deep learning models often use multi-layered ANNs like **convolutional neural networks (CNNs)** and **transformers**.
  - These networks learn intricate representations from data, enabling them to tackle complex tasks.

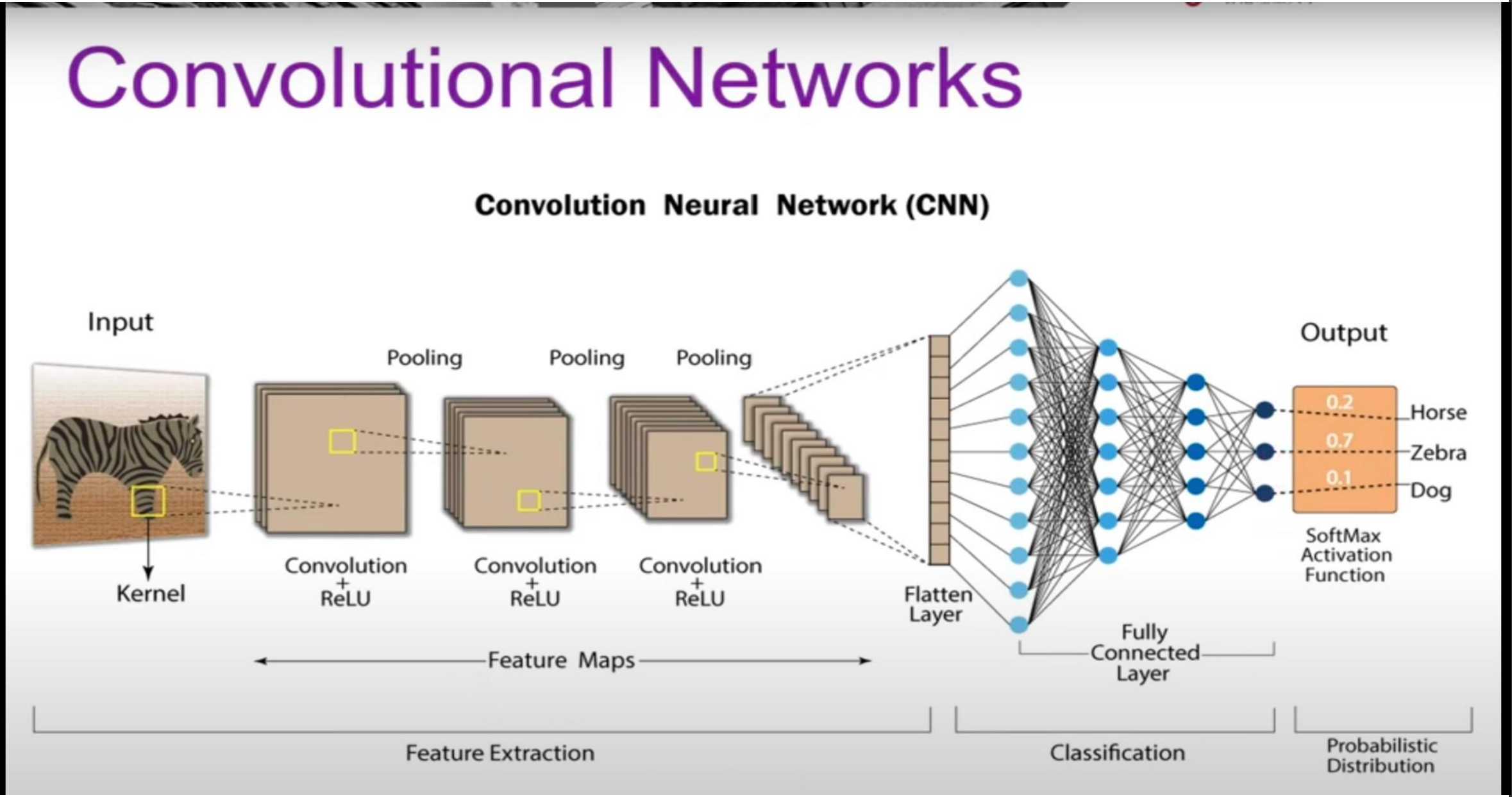


Convolutional Neural Network

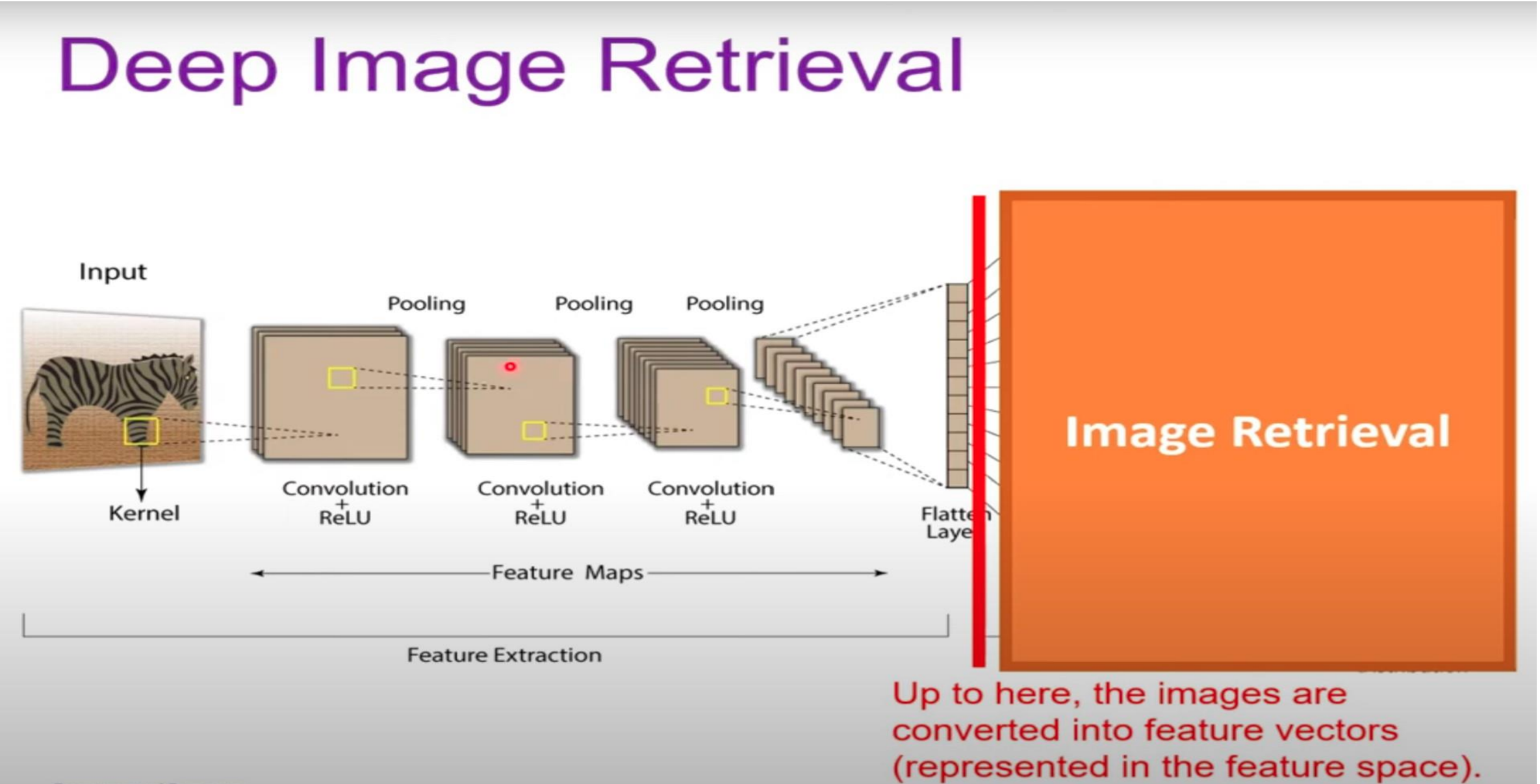
- CNNs are a class of deep neural networks specifically designed for processing structured grid-like data, such as images.
- They consist of multiple layers of convolutional, pooling, and fully connected layers.
- CNNs are adept at capturing spatial hierarchies of features in images, making them the backbone of many computer vision tasks.



General Representation of CNN

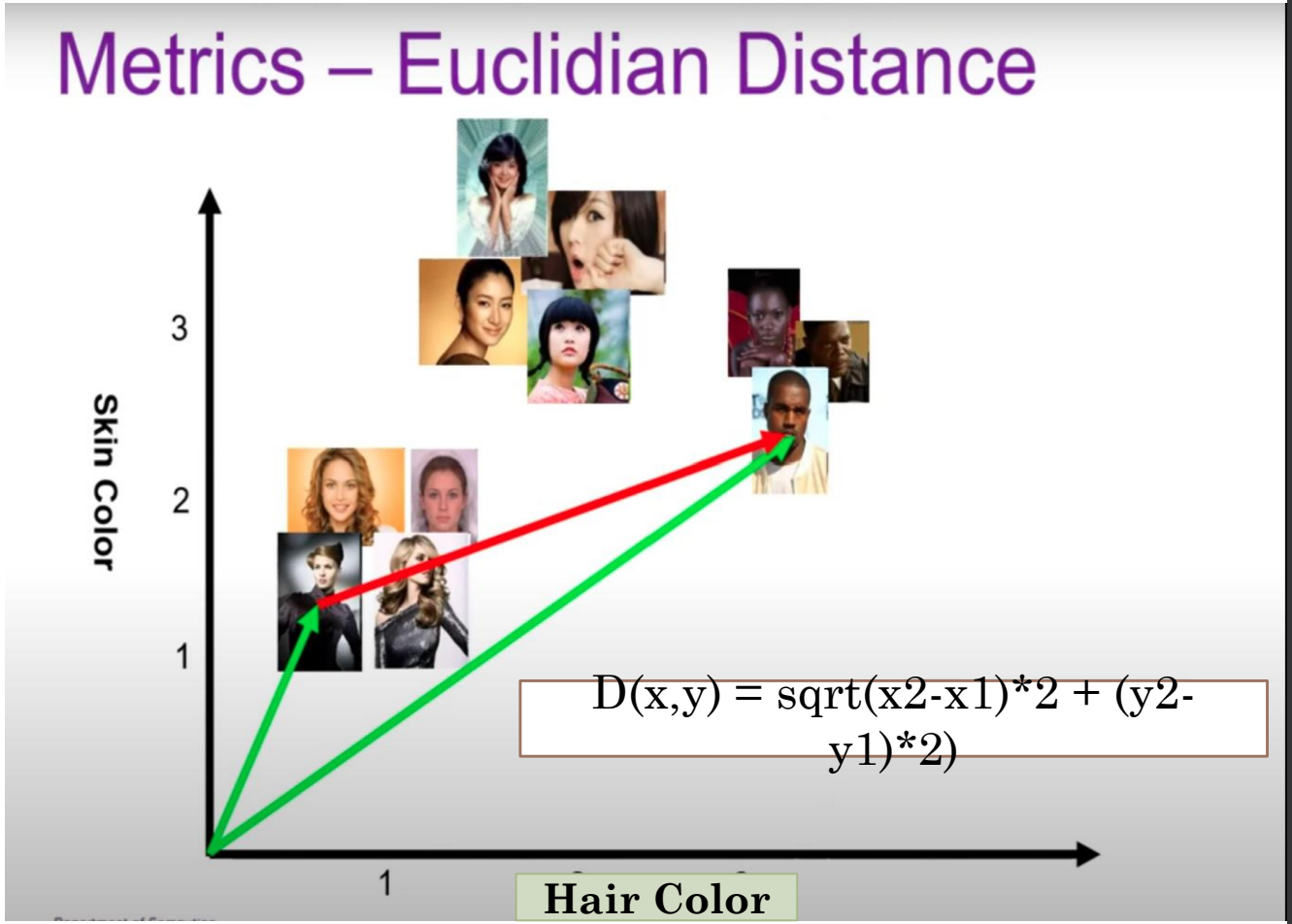


CNN for Image Retrieval



CNN FOR IMAGE RETRIEVAL

- As Long as the images are represented in the feature space.
- The search can be conducted by ranking images using similarity/distance metrics



- What is the best Way of using the Deep Features?
- Can we construct better features instead of using the raw feature maps?

deep features Uses

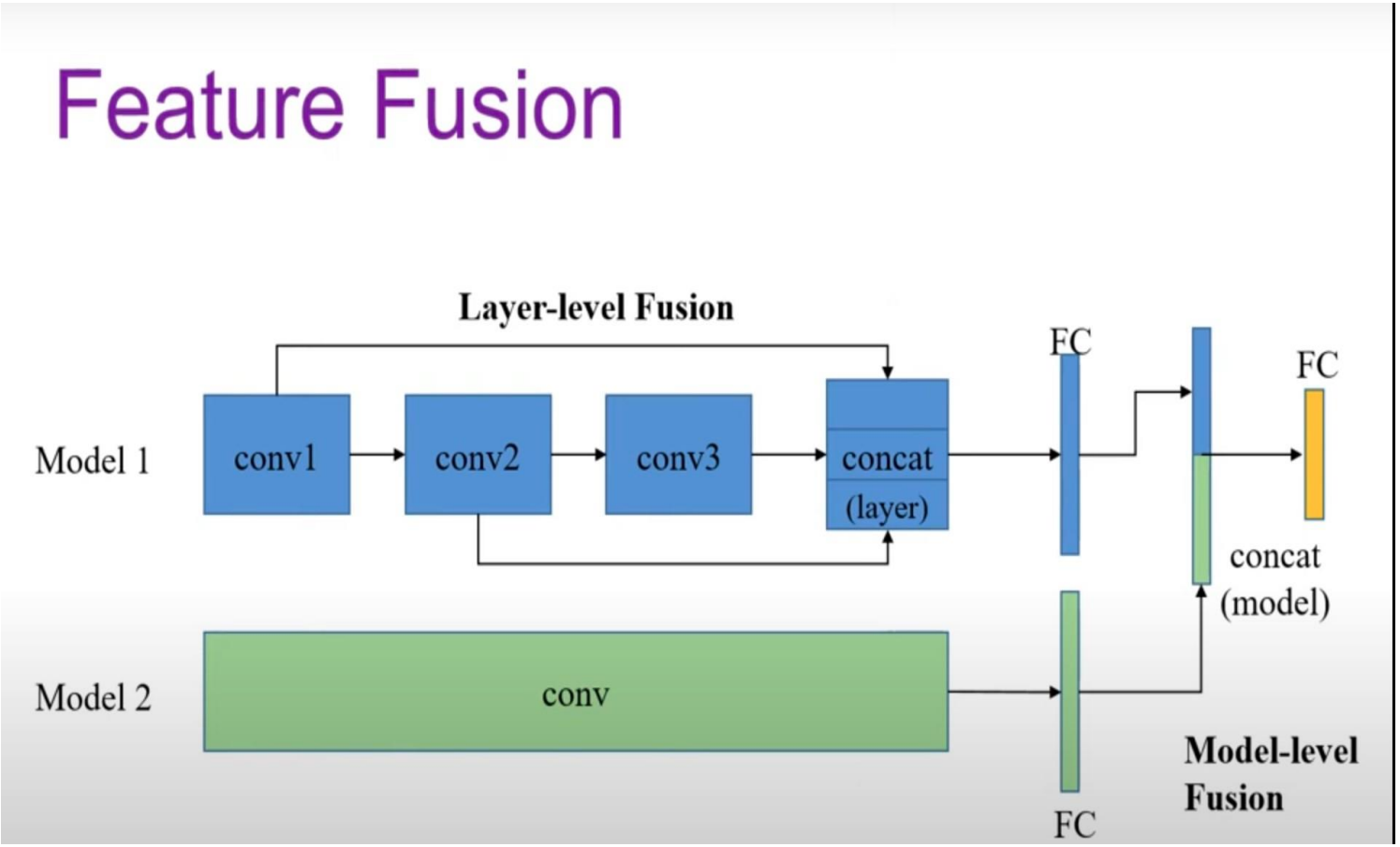
Feature Aggregation

In feature maps the spatial dimensions of the original images are “preserved”. We can thus **summarize** the features over the spatial dimensions for better representations of regions. This can be done by using different types of pooling algorithms.

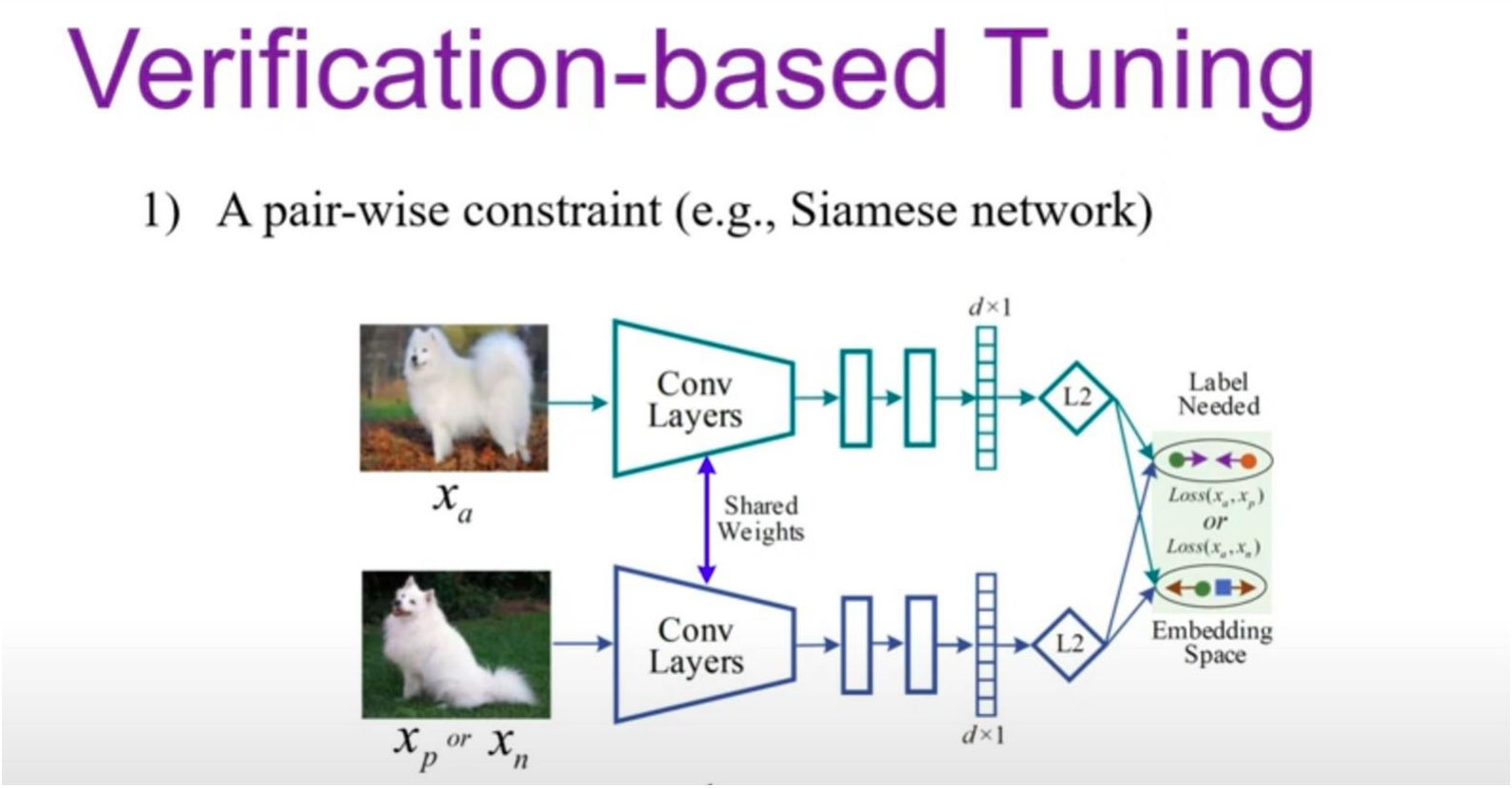
Sum/average Pooling **Max Pooling**

deep features Uses

- These methods are in fact called **Off-the-shell** methods.
- Because they don't change the parameters (weights) of the original CNNs.
- So there are fine-tuned-methods in which we update the parameters (weights) for better performance to address the domain shift.

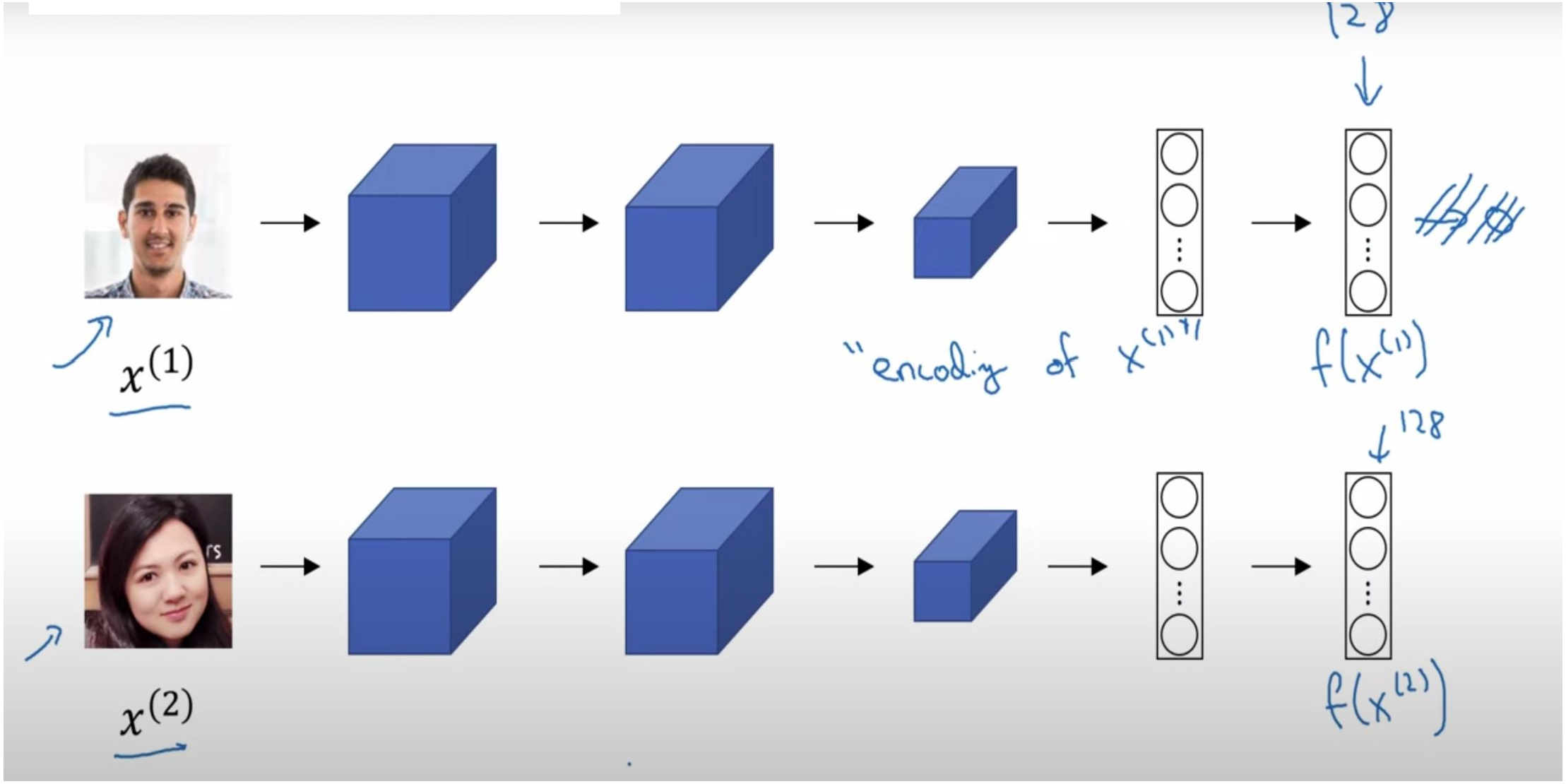


Verification based Tuning

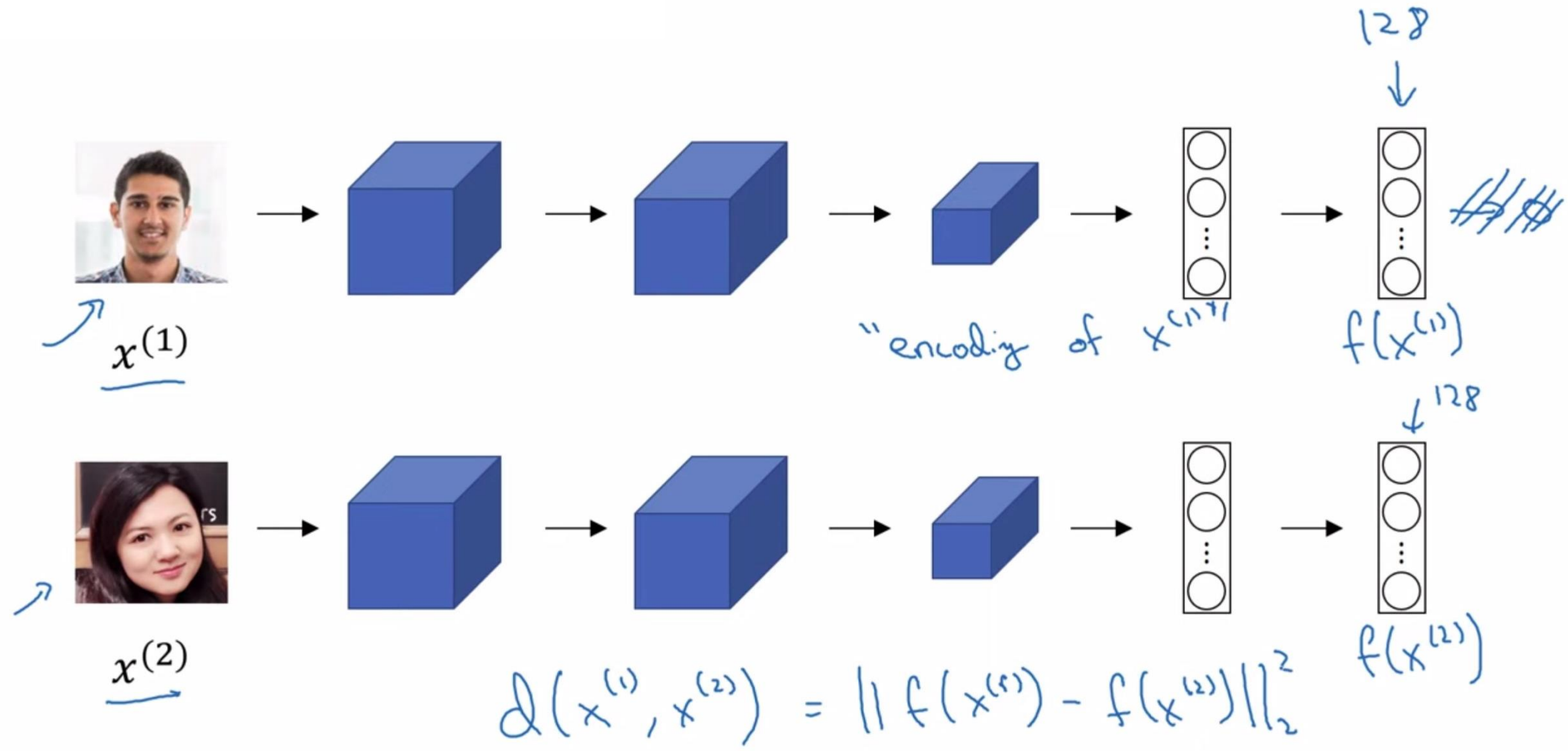


- The final output is a similarity score, indicating how similar or different the two input images are. This score can be used for predictions.

Siamese Network For Deep Face

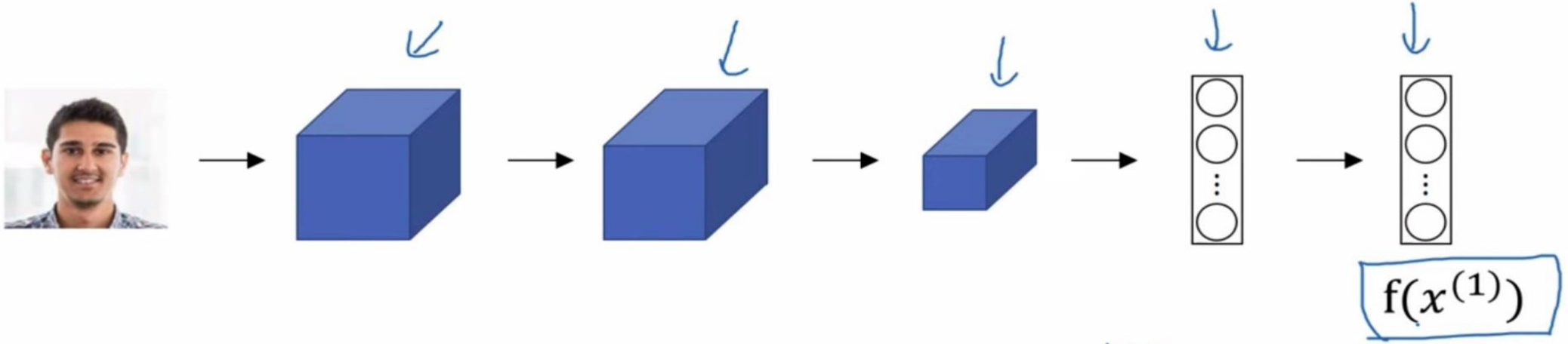


Siamese Network For Deep Face



Siamese Network For Deep Face

Goal of learning



Parameters of NN define an encoding $f(x^{(i)})$ 128

Learn parameters so that:

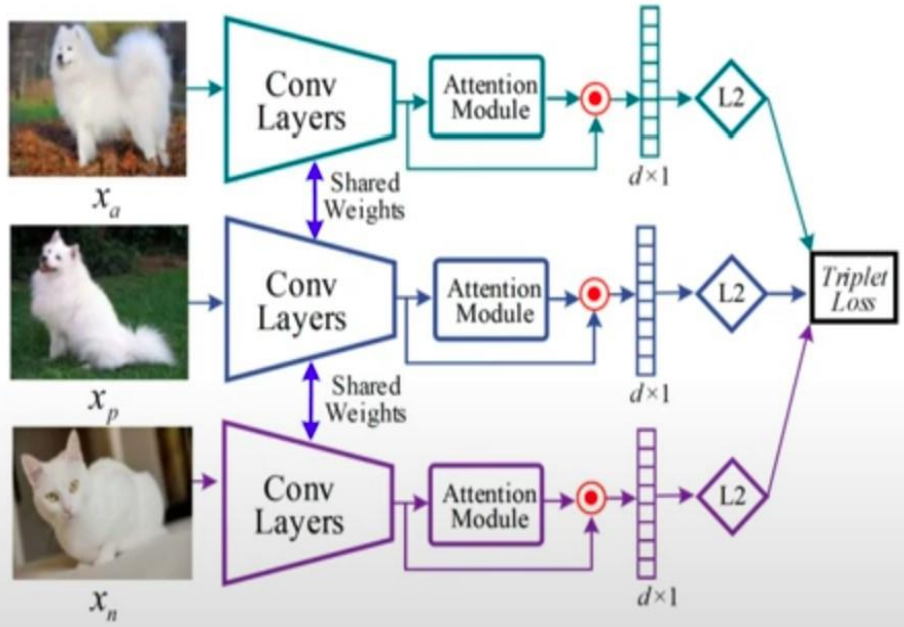
If $x^{(i)}, x^{(j)}$ are the same person, $\|f(x^{(i)}) - f(x^{(j)})\|^2$ is small.

If $x^{(i)}, x^{(j)}$ are different persons, $\|f(x^{(i)}) - f(x^{(j)})\|^2$ is large.

Verification based Tuning

Verification-based Tuning

2) A triplet constraint (e.g., triplet networks)



- The triplet consists of:
  - **Anchor**: The input sample.
  - **Positive Example**: An example with the same label as the anchor.
  - **Negative Example**: An example with a different label.

Triplet Network for face recognition

Learning Objective

Anchor — A

Positive — P

Anchor — A

Negative — N

Want: $\underbrace{\|f(A) - f(P)\|^2}_{d(A,P)} \leq \underbrace{\|f(A) - f(N)\|^2}_{d(A,N)}$

$\underbrace{\|f(A) - f(P)\|^2}_{0} - \underbrace{\|f(A) - f(N)\|^2}_{0} \leq \underline{0}$ $\quad f(\text{img}) = \vec{0}$

Triplet Network for face recognition

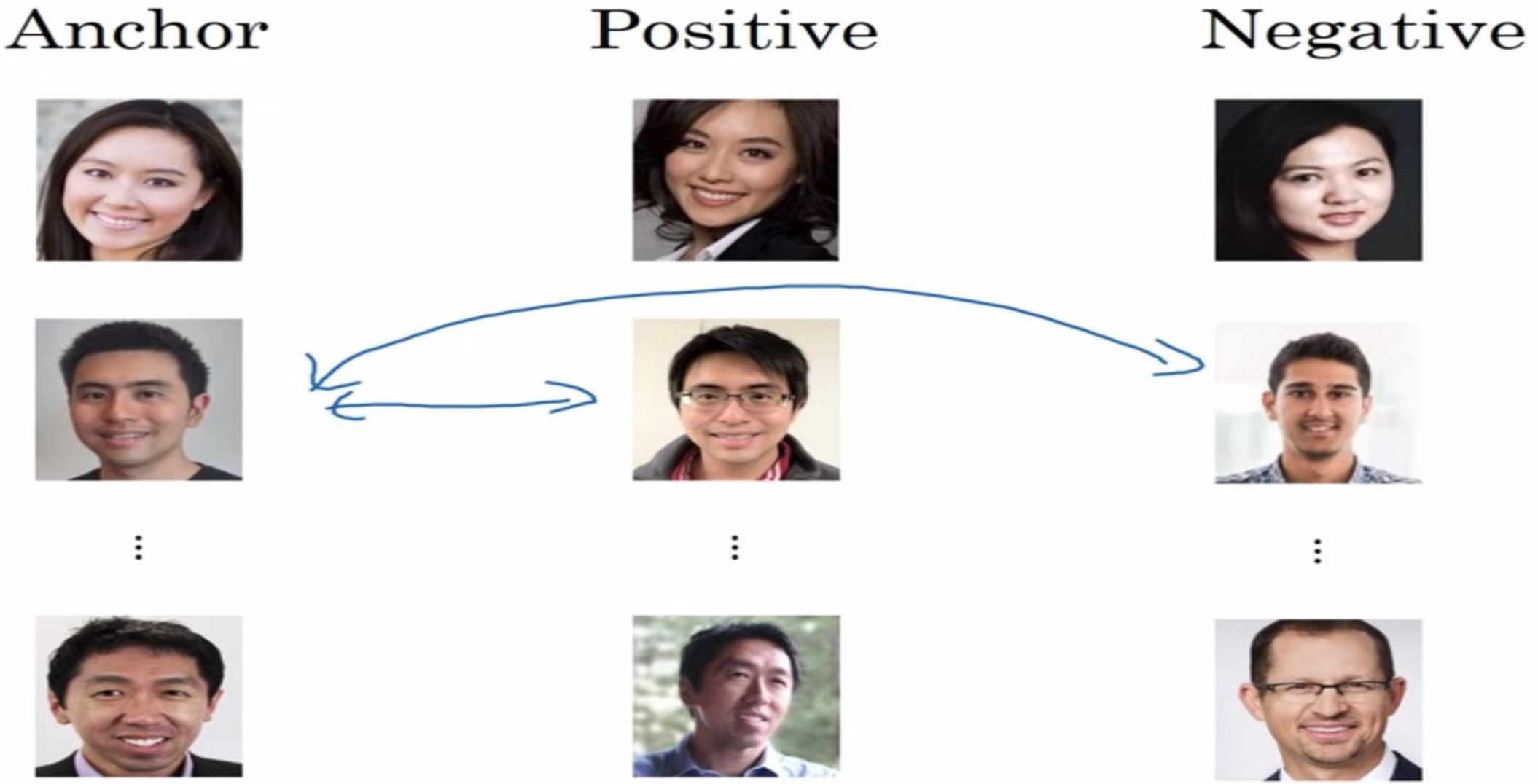
Loss function

Given 3 images A, P, N:

$$\mathcal{L}(A, P, N) = \max\left(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0\right)$$

$$J = \sum_{i=1}^{m} \mathcal{L}(A^{(i)}, P^{(i)}, N^{(i)})$$

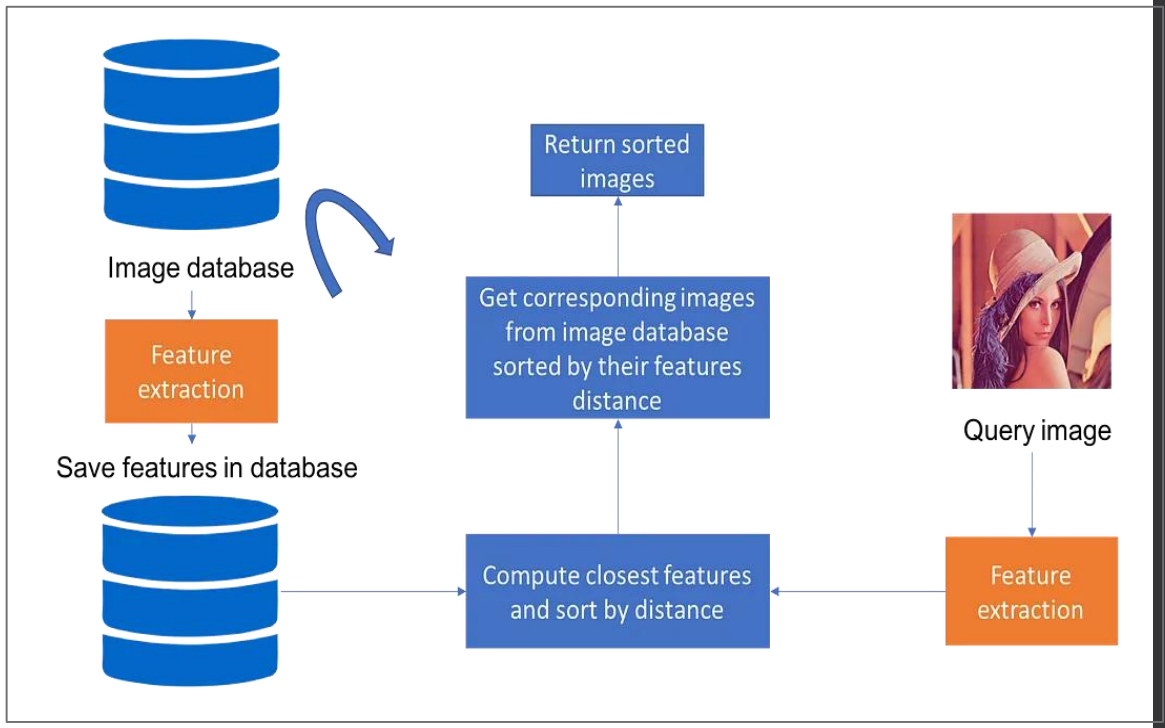
Triplet Network for face recognition



APPLICATIONS OF DEEP LEARNING IN IMAGE AND VIDEO RETRIEVAL

- **Content-Based Image Retrieval (CBIR)**:

- CBIR systems use visual features (such as color, texture, and shape) to retrieve similar images from a database.
- Deep learning models can learn powerful representations from raw pixel data, enabling more effective CBIR systems.
- Given a query image, these models can find visually similar images by comparing their learned features.

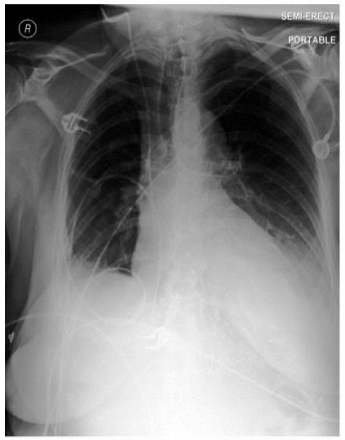


APPLICATIONS OF DEEP LEARNING IN IMAGE AND VIDEO RETRIEVAL

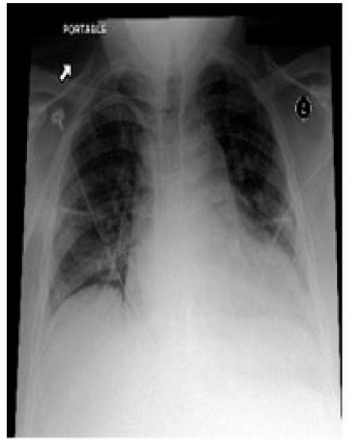
- **Object detection and recognition**:
- Deep learning models trained for object detection and recognition can facilitate image and video retrieval by accurately identifying specific objects or scenes within the media content.

APPLICATIONS OF DEEP LEARNING IN IMAGE AND VIDEO RETRIEVAL

- **Medical Imaging:**
- Deep learning can assist healthcare professionals in retrieving similar medical images for diagnosis, treatment planning, and research purposes.
- It helps in identifying patterns, anomalies, and relevant cases from large medical image databases.



(a) Test Image



(b) Least Similar Image

(c) Most Similar Image

APPLICATIONS OF DEEP LEARNING IN IMAGE AND VIDEO RETRIEVAL

- **Face Recognition:**
- Face recognition involves detecting faces, extracting features, and matching them to known faces using deep learning-based techniques.
- Deep learning enables automatic feature extraction and robust representation learning for accurate face recognition in various conditions.
- Face recognition in information retrieval systems relies on deep learning models to efficiently identify individuals and retrieve relevant information.