Deep Learning

Dr.Jawwad Ahmad Shamsi ,
National University of Computer and Emerging Sciences, Karachi Campus, Pakistan

February 26, 2024

1 Facial Recognition vs Facial Verificaiton

Face Verification This is a task that given an input <image, name/ID> Output wether the input image is of the claimed person or ID

Face Verification

Given an input image and a set of 'n' stored images, determine wether the input image matches with any of the stored image.

2 One Shot Learning

One shot learning refers to the phenomenon that for most of the cases, we will have a single image to learn or recognize a person or will be given just one example of that person's face.

Deep Learning algorithms doesn't work well if we have a single image or training example. For instance, we have a database of four pictures of students in a class. Now assuming that a student shows up and we want to recognize that someone shows up at the class and they want to be recognized. If this is a new person, then this should not match with existing dataset.

The challenge is that we only have one example of each person.

One approach we could try is to input the image of the person feed it to a COnvent and have it output a label with softmax with five outputs (four for existing persons in database and the fifth for a new person). This has two issues:

1. robustness decreases with more persons in the database
2. We need to retrain every time a new person is added in the system

We need a neural network which compares the two images and outputs the degree of difference between the two images such that if the two images are of different persons then the output is high.

so during recognition time if the degree of difference between them is less than some threshold (tau) then we would predict that these two pictures are of the same person and it is greater than tau you would predict that these are different persons

3 Facial Encodings and Siamese Network

[Net]

So the first thing we need is to create encodings of the facial images.

We will represent faces through encodings of 128 bytes. Figure 1 illustrates the concept of creating facial embeddings. So we have a DNN whcih applies convolution to an input face and creates encodings. The idea is that similar faces will have similar encodings and faces which are different from each other will have different encodings.

To assess the similarity of the two faces, we will use distance d between the two faces. Assuming that $x^{(1)}$ and $x^{(2)}$, represents encodings of the two faces, respectively,

if $x^{(i)}$ and $x^{(j)}$ are of different persons, then we want that distance between their encodings to be large. Comparatively, if $x^{(i)}$ and $x^{(j)}$ are of similar persons, then we want that distance between their encodings to be small (see figure 2).

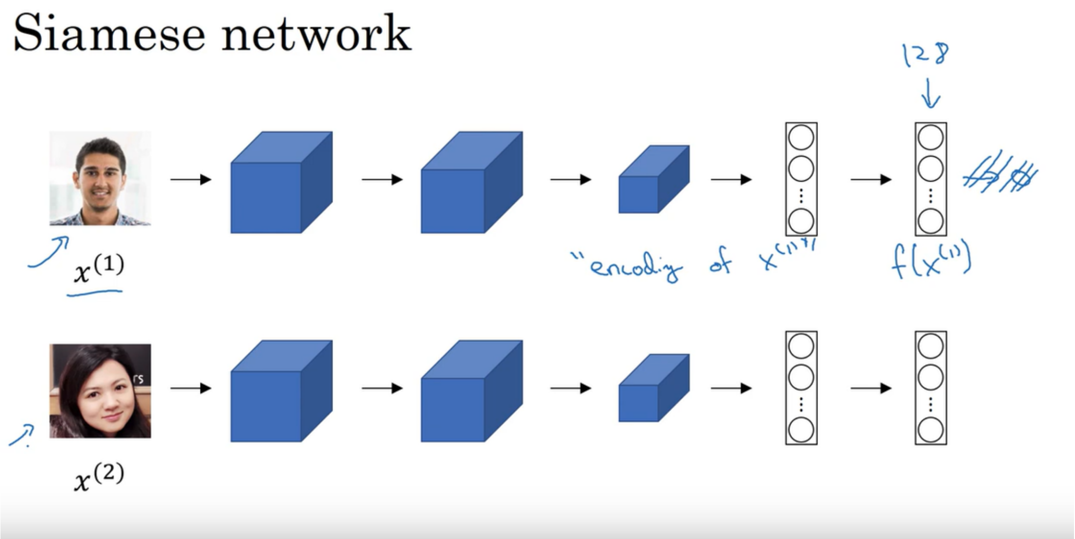
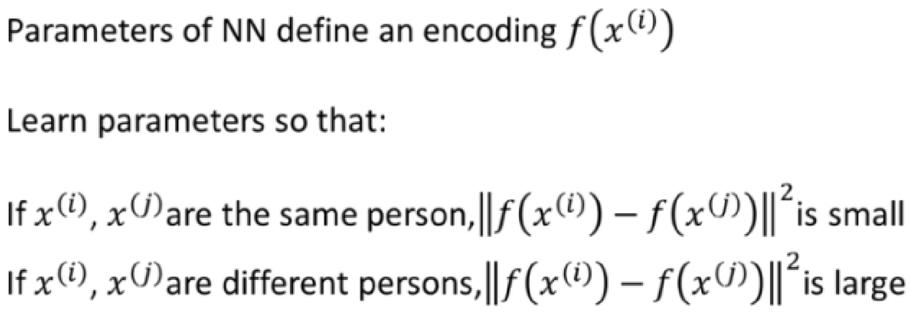


Figure 1: Encodings through Siamese Network
[Ng]



Parameters of NN define an encoding $f\left(x^{(i)}\right)$

Learn parameters so that:

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

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

Figure 2: Siamese Encoding
[Ng]

4 Triplet Loss

[Moi]

To train the siasme network, we use triplet loss function.

The goal here is as follows:

1. Two examples with similar faces should have similar embeddings vector
2. Two examples with dissimilar faces should have a large difference between the embeddings vector.

We compare images using a triplet:

1. Anchor (A)
2. Positive(P) : same as the Anchor
3. Negative (N) : Different than the anchor

Assuming that the function $d(i, j)$ denotes the distance (dissimilarity) between the two images i and j, we can assume that

$$d(A, P) <= d(A, N) \tag{1}$$

$$||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 <= 0 \tag{2}$$

To adjust the inequality, we can add α. However, we would like α to be significant. Note that α is a hyper parameter.

$$d(A, P) + \alpha = d(A, N) \tag{3}$$

$$||f(A) - f(P)||^2 + \alpha = ||f(A) - f(N)||^2 \tag{4}$$

So given 3 images A,P,N, we can define our loss function as follows:

$$L(A,P,N) = Max(||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \alpha, 0) \tag{5}$$

In the above equation if the first term is greater than 0, then it will be selected , otherwise, 0 will be selected.

4.1 Training

We can train this network using a series of m images. should be greater than one. Remember, this still holds the one shot leanring as we are only dealing with a single instance for inference. However, for training, we need to have multiple instances of a face (say m=10)

$$J = \sum_{i=1}^{m} L(A^i, P^i, N^i) \tag{6}$$

In order to select A,P, and N we do not select them randomly. we select them intelligently such that

$$d(A,P) \approx d(A,N) \tag{7}$$

This is because we would like this difference to be low for training so our network is able to train even on a minor difference.

References

[Moi] Olivier Moindrot. *Triplet Loss and Online Triplet Mining in TensorFlow.* URL: https://omoindrot.github.io/triplet-loss. (accessed: 01.01.2023).

[Net] Siasmee Network. *Siasme Network.* URL: https://datahacker.rs/one-shot-learning-with-siamese-neural-network/. (accessed: 01.01.2023).

[Ng] Andrew Ng. *Deep Learning Lectures on You Tube by Andrew Ng.* URL: https://www.youtube.com/@Deeplearningai. (accessed: 01.01.2023).