MLANN; Maximum Likelihood Approximate Nearest Neighbor in Real-time Image Recognition

ALI GHOLAMI
DEPARTMENT OF COMPUTER ENGINEERING & INFORMATION TECHNOLOGY
AMIRKABIR UNIVERSITY OF TECHNOLOGY
https://aligholamee.github.io
aligholami7596@gmail.com

***Abstract*—In this report, a brief overview of the *OpenFace* framework and the method behind it; *FaceNet* is reviewed for the task of *Image Recognition*.**

I. INTRODUCTION

AN evolutionary paper [1] caused the *Image Recognition* task to be more **accurate** and much more suitable for devices with lower **computational** power such as mobile devices. The mentioned paper has been implemented elaborately in the *OpenFace* framework. The method is reviewed to gain a better understanding of how *Image Recognition* methods can be implemented in practice.

II. FACENET PAPER GOALS

This paper supports the following three main goals for the task of *Image Recognition*:

- Face Verification – Is this the same person?
- Face Recognition – Who is this person?
- Face Clustering – Find common people among these faces?

III. THE IDEA

The idea is that, we can learn Euclidean embedding per image using *CNN* and the network is trained such that the squared L2 distances in the embedding space directly correspond to **face similarity**. The mentioned distance of two images illustrates whether two images are identical or not. In more details, a distance of 0.0 illustrates the *identity* of two images and a distance of 4.0 shows that two images are completely in a *opposite spectrum*. According to the results, a threshold of 1.1 can correctly classify the images. Thus; if the distance of two images is less than 1.1 it is representing the same person, otherwise they are different.

IV. FACENET'S METHOD

In order to reach the given idea, we have to consider using three images in every step of training:

1) **Anchor** image – The **actual** image we are learning the parameters with.

Advisor: Professor Mohammad Rahmati

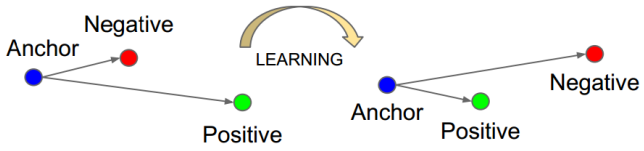

Fig. 1. The Triplet Loss minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.

2) **Positive** image – The image that is representing the **same** person as in the Anchor image.
3) **Negative** image – The image that is representing a completely **different** person.

We want to *minimize* the distance between the *Anchor* image and the *Positive* image. Also, we want the distance between the *Anchor* image and the *Negative* image to be *maximized*. Let $f(x)$ denote the embedding of image x into a d dimensional feature space R^d. The given method can be written as:

$$||f(A) - f(P)||^2 \leq ||f(A) - f(N)||^2$$

note that we have used the *L2* distance. As always, we have to propose a *loss* function. The loss function for the given formal representation is:

$$L(x_i^a,\ x_i^p,\ x_i^n) = \sum_{i=0}^{N}[||f(x_i^a) - f(x_i^p)||^2 - ||f(x_i^a) - f(x_i^n)||^2 + \alpha]$$

which x_i denotes the ith image in a set of N images. α is a margin term that allows the faces for **one** identity to live on a **manifold**, while still enforcing the distance and discrimination from other identities.

REFERENCES

[1] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.