CS419 Project Report

May 2022

1 Group Details

- Aditya Kudre 200070039
- Kaiwalya Joshi 20D070043
- Vaibhav Kumar 20D070087
- Pavan Bodke 200070014
- Rituraj Chaudhari 20D070065

2 Introduction

2.1 FaceNet

FaceNet is a CNN architecture widely used in face recognition. It compares two faces by first calculating the embedding vectors and then taking the euclidean distance between them. In this model we have a stack of layers that transforms the input image into a 128 dimensional vector. It is trained using the triplet loss. Let consider a triplet of 3 images anchor,positive and negative. Positive image is the image that belongs to the same class as the anchor while negative image is the image that belongs to a different class than the anchor. Here 2 images are the same class if they belong to the same person. Triplet loss is defines as

$$[||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha]_+ \tag{1}$$

The stack of layers or the embedding network that we defined in the previous paragraph can be of different types like Zeiler and Fergus,GoogleNet and VGGNet. In our project we have majorly analysed Zeiler and Fergus architecture. Here is a brief description of this architecture.

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

Figure 1: Zeiler and Fergus based Architecture

Now there are different ways to choose the triplets of face images from a dataset. In a most basic way they can be randomly selected and we can also mine them using hard or semi-hard mining. In hard mining we select positive image such that it is farthest away from the anchor image and negative image is selected such that it is closest to the anchor image. While in semi-hard mining triplets are selected such that

$$||f(x_i^a) - f(x_i^p)||_2^2 < ||f(x_i^a) - f(x_i^n)||_2^2 \tag{2}$$

2.2 Haar cascades

Haar Cascades is a way to detect general objects in an image. It is particularly used to detect faces in an image. We have implemented Haar cascade to crop faces from an image.

3 Experiments

We have implemented the FaceNet model with Zeiler and Fergus embeddings and semi-hard triplet mining. This is the graph that shows the variation of triplet loss over 10 epochs with $\alpha = 1$.

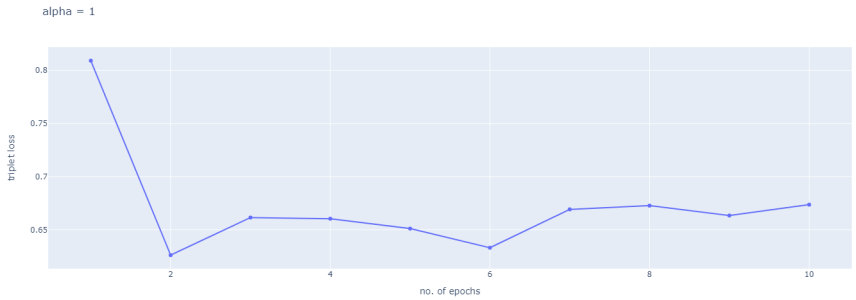

Figure 2: Variation of loss over 10 epochs for $\alpha = 1$

We also tried this model with different values of α like $\alpha = 0.2$ and $\alpha = 0.5$. Here are the graphs showing variation of loss over 5 epochs.

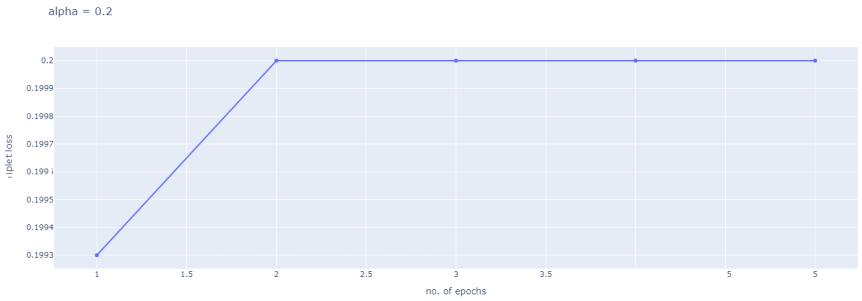

Figure 3: Variation of loss over 10 epochs for $\alpha = 0.2$

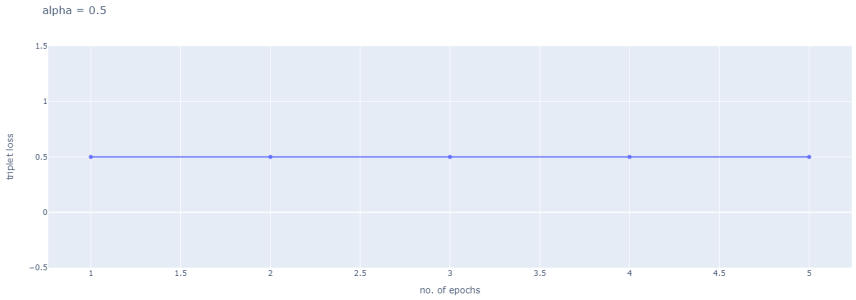

Figure 4: Variation of loss over 10 epochs for $\alpha = 0.5$

4 Links to Google Colab

- FaceNet Code : Link
- Haar Cascade experiment : Link