Evaluating models for Visual Similarity
Applied Deep Learning

Karan Sindwani(ks3631), Lisa Sarah Thomas(lt2709)

1. Introduction

Image/visual similarity is the measure of how identical two images are. In other words, it quantifies the degree of similarity between intensity patterns in two images. The applications of image similarity are manifold. The use cases span from visual search to detecting duplicate products on e-commerce websites.

2. Model

As a baseline, we used the **VGG16** pre-trained model. For each image in the training set, activations from the last layer of the VGG16 were extracted as features. Similarly, for each test image. The main idea behind extracting features is to get a vector representation/embedding for the image. Then, to compute the similarity between two images, euclidean distance is used as the distance metric. The image embedding with the smallest euclidean distance to the test image embedding is reported as most identical to the given test image.

The next model we evaluated is a deep neural network with **triplet loss**. The loss function takes as input 3 images: query/anchor image, positive image and negative image. The objective of the loss function is to learn an embedding function that assigns smaller distance to similar images. Figure 1 shows the mathematical representation of the same. ‘f’ is the embedding function to map image to vector, pi is the query image, pi+ is the positive image, pi- is the negative image, r is the similarity distance between 2 images.

$$
\begin{aligned}
&D(f(p_i), f(p_i^+)) < D(f(p_i), f(p_i^-)), \\
&\forall p_i, p_i^+, p_i^- \text{ such that } r(p_i, p_i^+) > r(p_i, p_i^-)
\end{aligned}
$$

Figure 1

As seen in the model architecture, the 3 input images are passed through separate deep neural networks. But, these networks share the same weights. The output of the network is the embedding representation of the input images.

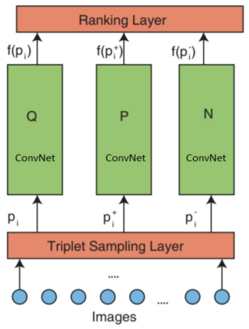

Fig 2. Model architecture

3. Dataset

The models we evaluated were trained on images from Labeled Faces in the Wild (LFW) dataset. LFW is a database of face photographs consisting of 13,233 images of 5,749 people. 1,680 of the people pictured have two or more distinct photos in the dataset. For the purpose of our experiments, we selected only those people who have at least 10 distincts photos. Our final dataset comprises of 1430 images of 143 people (10 images per person). The dataset has been further split into train-test (80-20) set to have a common test set while evaluating the models.

4. Evaluation

For evaluating the models, we have curated a separate test set. For a given test image, we retrieve 10 images from the train set that have the smallest euclidean distance. If the label of the test image appears in the 10 nearest images, it is considered a hit. The accuracy is then computed as the percentage of hits.

Model	**Accuracy Score**
Baseline model - VGG16	40.2%
Triplet loss - Resnet 50	59.3%

Youtube Link : https://youtu.be/8hkqOmk9zmE