Deep face Recognition

Summary

1. Propose a procedure to create large-scale face dataset while minimizing annotation effort.
2. Investigate CNN architecture for face verification and identification.

Dataset Collection

1. Dataset collection is a multi-step process emphasizing scale and data purity (precision)
2. Obtain names → collect images using search engine
   ↓
   improve data purity with a learned filter.
   ↓
   duplicate removal
   ↓
   manual filtering aided by a learned classifier.

Training

1. Training is a two-step process:
   - pretraining the network as a classifier
   - learning a projection operator with "triplet loss".
2. The triplet-loss training aims at learning an embedding of the input img such that comparing different input images by comparing their embedding is meaningful.
3. Given: $\phi(l_t) \in R^D$ is the output of the CNN
   - $x_t = W' \frac{\phi(l_t)}{\|\phi(l_t)\|_2}$
   - $W' \in R^{L \times D}$
   - x_t is l_2-normalized and projected version of $\phi(l_t)$
4. W' is trained to minimized:

$$E(W') = \sum_{(a,p,n) \in T} \max\{0, \alpha - \|x_a - x_n\|_2^2 + \|x_a - x_p\|_2^2\}, \quad x_i = W' \frac{\phi(l_i)}{\|\phi(l_i)\|_2}$$

5. α is a fixed scalar representing the learning margin
6. $\mathcal{T}$ is a collection of triplets.
7. A triplet (a, p, n) contains an anchor face image a, a positive $p \neq a$ and negative n examples of the anchor's identity.
8. The triplet (a, p, n) is chosen by extending each pair (a, p) to a triplet (a, p, n) by sampling the image n at random, but only b/t the ones that violate the triplet loss margin.
9. Face verification is to tell whether two images have the same identity or not.
10. This is obtained by testing whether the l_2 distance in the emb. space b/t the two images is smaller than a threshold τ.
11. This threshold τ is chosen to maximize the verification accuracy, rate of correctly classified pairs, on validation data.

Experiments

1. They also use Equal Error Rate (EER) as an evaluation metric, defined as the error rate at the ROC operating point where false positive and false negative rates are equal.
2. The advantage of this metric is that it is independent from the distance threshold τ.
3. Given a face image l, four 224×224 pixel patches are cropped from the four corners and the center with horizontal flip. The feature vector from these are averaged.
4. They train model with datasets obtained at different stages in the dataset curation process, to see what effects do the processing steps in the curation pipeline have?

Misc

1. The input to all network is a face image of size 224 × 224 with the average face image, computed from the training set, subtracted → this is critical for the stability of the optimization algo.

Questions

1. The triplet loss leads to embedding where the emb. of positive is close to emb. of anchor and the emb. of negative is far from emb. of anchor.

2. What does this sentence mean: "K face descriptors are obtained for each video by ordering the faces by their facial landmark confidence score"?