FaceNet: A Unified Embedding for Face Recognition and Clustering

Florian Schroff[1], Dmitry Kalenichenko[1], James Philbin[1] ({fschroff, dkalenichenko, jphilbin}@google.com)
[1]Google Inc.

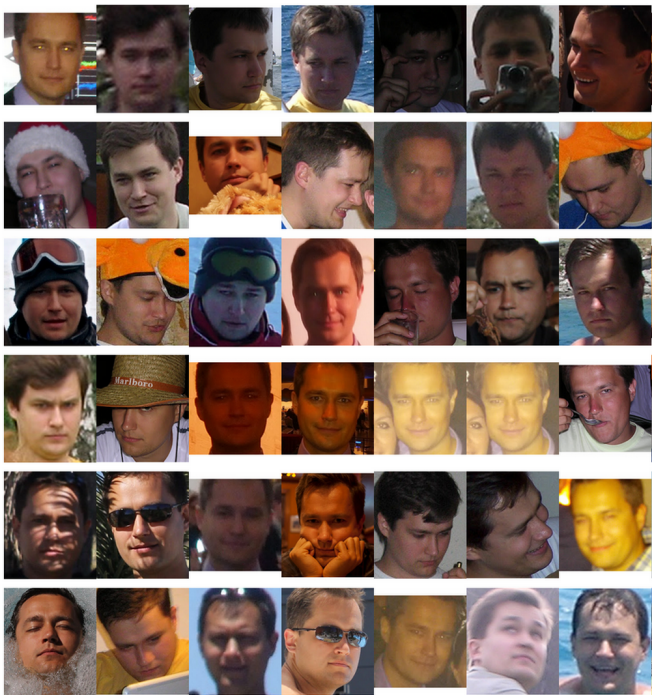

Figure 1: **Face Clustering.** Shown is an exemplar cluster for one user. All these images in the users personal photo collection were clustered together.

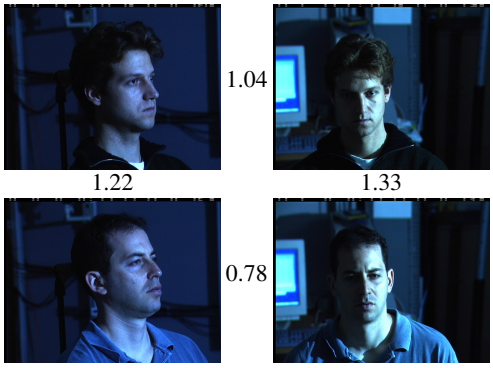

Figure 2: **Illumination and Pose invariance.** This figure shows the output distances of FaceNet between pairs of faces of the same and a different person in different pose and illumination combinations. A distance of 0.0 means the faces are identical, 4.0 corresponds to the opposite spectrum, two different identities. You can see that a threshold of 1.1 would classify every pair correctly.

Overview

Despite significant recent advances in the field of face recognition [1, 2], implementing face verification and recognition efficiently at scale presents serious challenges to current approaches. In this paper we present a system, called FaceNet, that directly learns a mapping from face images to a compact Euclidean space where distances directly correspond to a measure of face similarity. Once this space has been produced, tasks such as face recognition, verification and clustering can be easily implemented using standard techniques with FaceNet embeddings as feature vectors.

Figure 1 shows one cluster in a users personal photo collection. It is a clear showcase of the incredible invariance to occlusion, lighting, pose and even age. See Figure 2 for another illustration of its robustness to lighting and pose.

Figure 3: **Model structure.** Our network consists of a batch input layer and a deep CNN followed by L_2 normalization and the triplet loss.

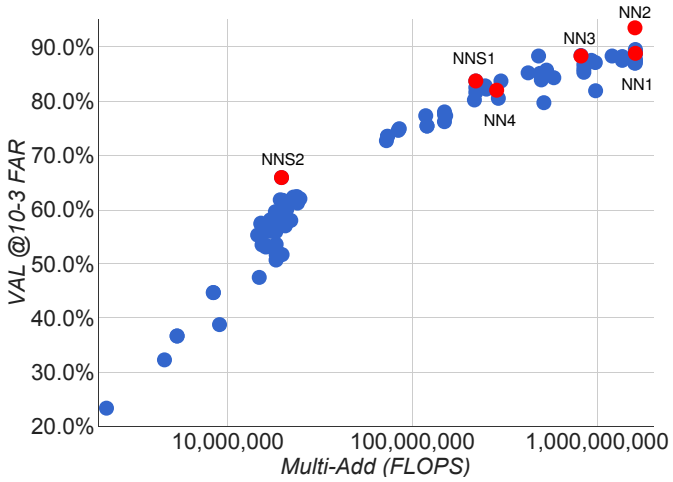

Figure 4: **FLOPS vs Accuracy trade-off.** Shown is the trade-off between FLOPS and accuracy for a wide range of different model sizes and architectures. Highlighted are the four models that we focus on in our experiments.

Our method uses a deep convolutional network trained to directly optimize the embedding itself, rather than an intermediate bottleneck layer as in previous deep learning approaches (see Figure 3). To train, we use triplets of roughly aligned matching / non-matching face patches generated using a novel online triplet mining method. We learn an embedding $f(x)$, from an image x into a feature space $\mathbb{R}^d$, such that the squared distance between *all* faces of the same identity is small, whereas the squared distance between a pair of face images from different identities is large.

The loss that is being minimized is then $L =$

$$\sum_i^N \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+ . \qquad (1)$$

α is a margin that is enforced between positive and negative pairs. Superscript a, p, n denote the anchor, positive and negative, respectively. The summation is over the set of all possible triplets in the training set.

The benefit of our approach is much greater representational efficiency: we achieve state-of-the-art face recognition performance using only 128-bytes per face.

We evaluate these architectures on a wide range of hyperparameters, leading to different trade-offs of computation and accuracy, see Figure 4.

On the widely used Labeled Faces in the Wild (LFW) dataset, our system achieves a new record accuracy of **99.63%**. On YouTube Faces DB it achieves **95.12%**. Our system cuts the error rate in comparison to the best published result [1] by 30% on both datasets.

[1] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. *CoRR*, abs/1412.1265, 2014.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conf. on CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.