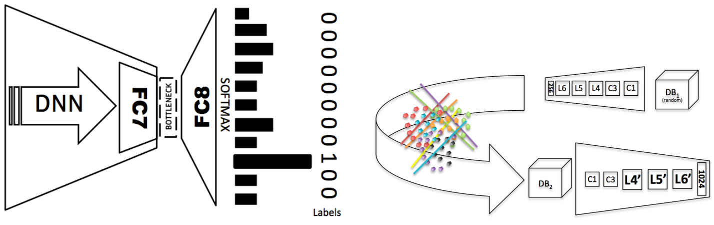
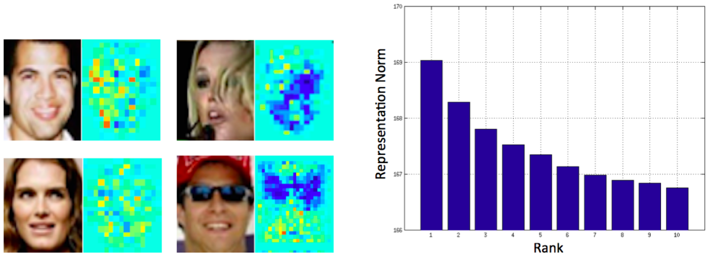
Web-Scale Training for Face Identification

Yaniv Taigman[1], Ming Yang[1], Marc'Aurelio Ranzato[1], Lior Wolf[2]
[1]Facebook AI Research, Menlo Park, California. [2]Tel Aviv University, Tel Aviv, Israel.



(a) (b)

Figure 1: (a) **The bottleneck.** The representation layer splits the network between the part that converts the input into a *generic* face descriptor and the part that performs linear classification to *specific* K classes. (b) **The bootstrapping method.** An initial 256D-compressed representation trained on DB_1 is used to find the semantically-nearest identities of randomly picked 100 seeds, in a large pool of pre-trained hyperplanes. The union of all 100 groups of selected identities define the bootstrapped dataset DB_2. A larger capacity network with enlarged locally-connected layers and a 1024D representation is then trained.



(a) (b)

Figure 2: (a) **The Representation Norm.** Examples of L6 activations for various faces. In each pair the original image is compared to the sum of the channels of its L6 activations: (left) pairs depicting good quality images. (right) examples of poor quality images (occluded and/or misaligned). Bluer is lower, red is higher. (b) **Retrieval rank vs. mean representation norm** on our internal validation set. Misclassified probes (rank>1) tend to have a lower norm than correctly matched probes (rank=1).

We study face recognition and show that three distinct properties have surprising effects on the transferability of deep convolutional networks (CNN): (1) The bottleneck of the network serves as an important transfer learning regularizer, and (2) in contrast to the common wisdom, performance saturation may exist in CNN's (as the number of training samples grows); we propose a solution for alleviating this by replacing the naive random subsampling of the training set with a bootstrapping process. Moreover, (3) we find a link between the representation norm and the ability to discriminate in a target domain, which sheds lights on how such networks represent faces. Based on these discoveries, we are able to greatly improve face recognition accuracy on the widely used LFW benchmark, both in the verification (1:1) and identification (1:N) protocols, and directly compare, for the first time, with the state of the art Commercially-Off-The-Shelf system and show a sizable leap in performance.

Face identification is a recognition task of great practical interest for which (i) much larger labeled datasets exist, containing billions of images; (ii) the number of classes can reach tens of millions or more; and (iii) complex features are necessary in order to encode subtle differences between subjects, while maintaining invariance to factors such as pose, illumination, and aging.

Our contribution is not only to redefine the state of the art on a public benchmark using an improved system, but also: (i) we study the role of the bottleneck as a transfer learning regularizer and (ii) we propose a new way to utilize large datasets by replacing the standard random sub sampling procedure with a bootstrapping procedure; (iii) we discover a three-way link between the representation norm, the image quality, and the classification confidence.

Transferring with Bottleneck Starting with an initial face representation, which was trained similarly to DeepFace [4], we exploit discoveries in this network to scale the training up considerably. We show that the dimensionality of the last fully-connected layers (F7 & F8) critically affects the balance between generality and specificity. As illustrated in Fig 1(a), in a K-way multiclass network with binary targets, the *classification* layer (F8) is a collection of K linear (dependent) classifiers. By compressing the preceding *representation* layer (F7) through a lower rank weight matrix, we reduce the ability of the network to encode training-set specific information in this layer, thereby shifting much of the specificity to the subsequent classification layer (F8).

Semantic Bootstrapping A conjecture is made in [3] that "results can be improved simply by waiting for faster GPUs and bigger datasets to become available". Our findings reveal that this holds only to a certain degree. We find that using the standard Stochastic Gradient Descent (SGD) and back-propagation leads to performance saturation in the target domain when the training set size in the source domain grows beyond a certain point. This holds even when we change the architecture of the network, by either adding more layers and/or increasing their capacity. As illustrated in Fig 1(b), by judiciously selecting samples as opposed to picking them at random, we were able to improve performance further.

Representation Norm Since image disruptions lead to localized L6 inactivity, and since F7 is a linear projection of L6 followed by a threshold, these disruptions lead to a reduced norm of the representation vector F7. This can be seen in Fig. 2(a). Moreover, we show that there exist a link between the representation norm and the predicted retrieval rank (Fig. 2(b)), which can be used to understand weak retrievals.

Results Our best method lowers the Commercially-Off-The-Shelf state of the art [1, 2] miss rate on the LFW closed set protocol by 57% (82% rank-1), and by 45%, at the same precision level, on the open set protocol. The error in the verification protocol is reduced by 38% with respect to the initial baseline system, and stands on 98.37%.

[1] Lacey Best-Rowden, Hu Han, Charles Otto, Brendan Klare, and Anil K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. IEEE Transactions on Information Forensics and Security, 2014.

[2] Patrick J. Grother, George W. Quinn, and P. Jonathon Phillips. Report on the evaluation of 2D still-image face recognition algorithms. NIST Multiple Biometrics Evaluation (MBE), 2010.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[4] 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 Computer Vision and Pattern Recognition (CVPR)*, 2014.

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