

Blinding with Linear Clustering Removal

Samuel Wehrli

January 22, 2021

Abstract

This work investigates discriminatory dimensions in face recognition algorithms. A simple blinding procedure is proposed to remove information in the embedding space related to clustering with respect to the discriminatory dimension. The procedure is a linear operation on the embedding vectors which removes the cluster separation. After outlining the method, cluster visualization, classification performance with regard to the clusters and face recognition rates are investigated before and after the blinding procedure.

1 The math behind

In the following we look at the discriminatory dimension of race. We work with the commonly used racial faces in-the-wild (RFW) data set which has the four ethnic labels Caucasian, African, Asian and Indian. Face recognition algorithms work with embedding spaces. They map images of persons into a vector space with dimension N_e . In the following I consider a VGG2 model where $N_e = 128$. The mapping is optimized such that pictures of the same person are close to each other in the embedding space. The proposed procedure takes the embeddings x_i as an input (i denotes the sample). Each sample corresponds to an ethnic class k . The goal is to remove those directions of the embedding space, which separate the ethnic clusters. It is mathematically equivalent to a unitary rotation in the embedding space such that some of the features capture the cluster separation and a subsequent removal of these features. This operation can be carried through the following steps. First we define the centers of each cluster simply by averaging

$$\bar{x}_k = \frac{1}{n_k} \sum_{i \in C_k} x_i, \quad (1)$$

where C_k is the set of embedding vectors associated with cluster k and n_k is the corresponding size. Following a one-vs-rest (OvR) approach, the normalized separational directions of each cluster center \bar{x}_k to the other centers is given by the vectors

$$u_k = \frac{v_k}{\|v_k\|} \quad \text{with} \quad v_k = \bar{x}_k - \frac{1}{K-1} \sum_{k' \neq k} \bar{x}_{k'}, \quad (2)$$

where K is the number of clusters. Fig. 1 shows the projections onto the vectors u_k . As it may be expected by the construction, each direction nicely separates the corresponding cluster from the others. These means, that different ethnics groups are literally located in different corners of the embedding space. The goal of the remaining step is to remove this separation by simply projection out the corresponding directions u_k . By construction, the vectors u_k are not linearly independent, but span a subspace of rank $K-1$. This can be verified by applying a singular value decomposition (SVD). SVD also the corresponding basis vectors b_j where $j = 1 \dots K-1$. The final step is to remove the separational directions with

$$x_i^b = x_i - \sum_{j=1}^{K-1} (x_i \cdot b_j) b_j, \quad (3)$$

where $(x_i \cdot b_j)$ is the dot product. Eq. (3) yields new embedding vectors with the same shape as the original ones. The upper index b stands for *blinded* inspired by the fact that some information with regard to the discriminatory dimension has been removed. Note that new embeddings depend linearly on the original ones.

2 Awareness

Awareness is the ability of the model to discriminate between different clusters of the discriminatory dimensions, being the ethnic label in present case. This ability obviously depends on the algorithms. Here I benchmark the performance to predict the ethnic labels by the *aware* embeddings x_i and the *blinded* ones x_i^b . Both embeddings have

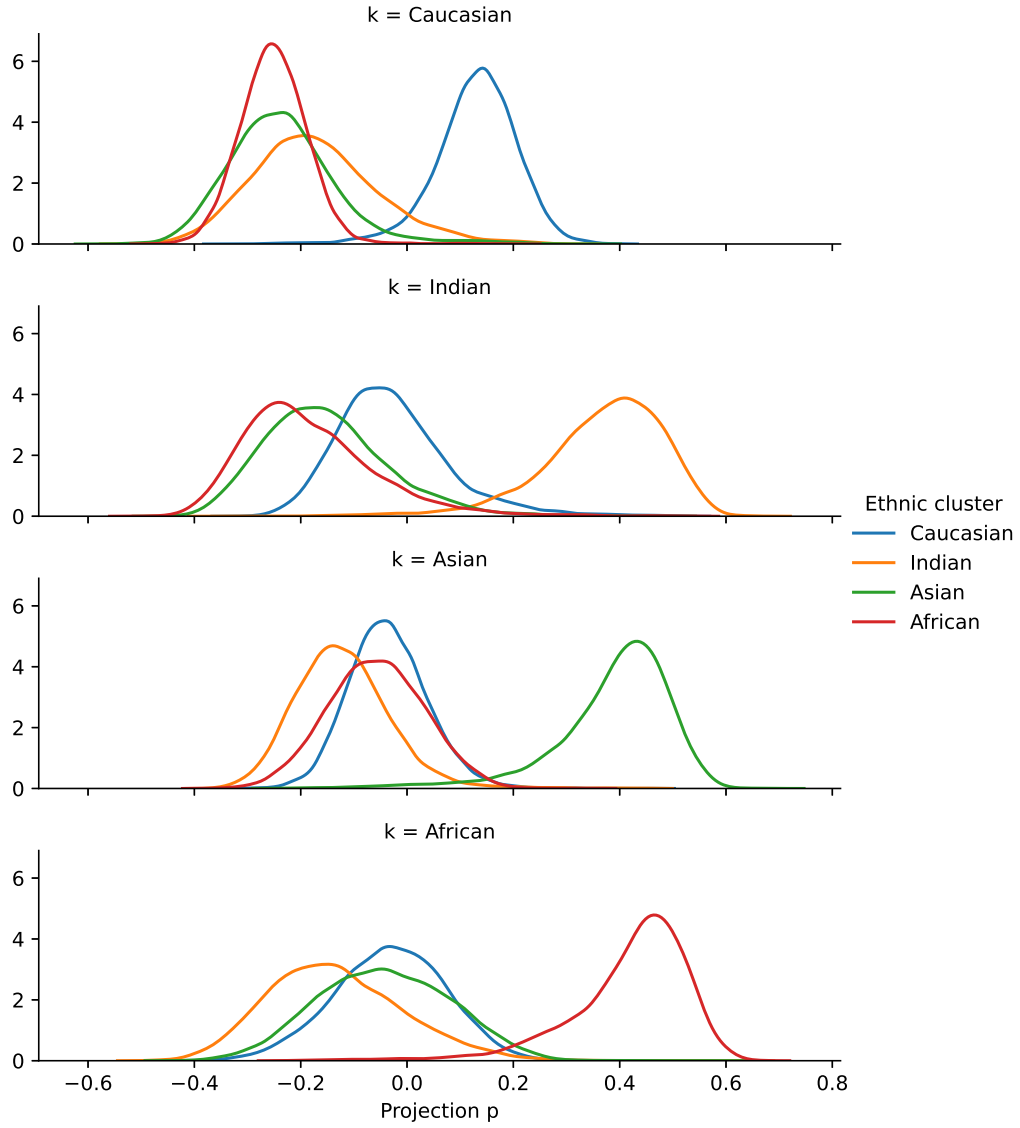


Figure 1: Distribution of projections $(x_i \cdot u_k)$. Each direction u_k separates the corresponding cluster from the others.

Model	aware	blinded
Logistic regression	96%	21%
Linear SVM	96%	26%
Nearest neighbor	92%	57%
5 Nearest neighbor	94%	62%
NN with 1 hidden layer (100 nodes), relu	96%	70%
NN with 2 hidden layer (100 nodes each), relu	96%	85%

Table 1: Subset accuracy of various classifiers predicting the ethnic labels based on *aware* and *blinded* emdeddings. A train/test split of two thirds/one third was used.

been split into 66.6% training data and 33.3% test data. The accuracy of different classifiers is shown in Tab. 1. Not surprisingly, linear classifiers are unable to predict the race for the blinded embeddings. Nearest neighbor approaches still work reasonably. More advanced non-linear classifiers such as neural networks perform well. The clustering displayed by the corresponding t-SNE plots in Fig. 2 is in line with these findings. In the blinded case, clusters can’t be separated by a single straight line. However, the data still displays groups defined by ethnical labels. Interestingly, there are clear differences between the groups. Africans are grouped in the center, Caucasian encircle the cloud and Indians/Asian are scattered inbetween.

3 Cluster scores

Cluster scores give a measure of clustering. They are calculated for both embeddings in Tab. 2. The Silhoutte cluster score gives a measure between 0 and 1 indicating how well the data is clustered. The Silhoutte cluster score of the aware embeddings is only 0.063 - basically indicating the absence of clustering although Fig. 1 and Fig. 2 show nice clustering for the aware embeddings. This counter-intuitive finding is due to the high dimensionality. Both figures show projections into lower dimensions and therefore reflect only a marginal part of the information. This is confirmed by the fact that the total variance of the blinded embeddings is still 84% of the original variance.

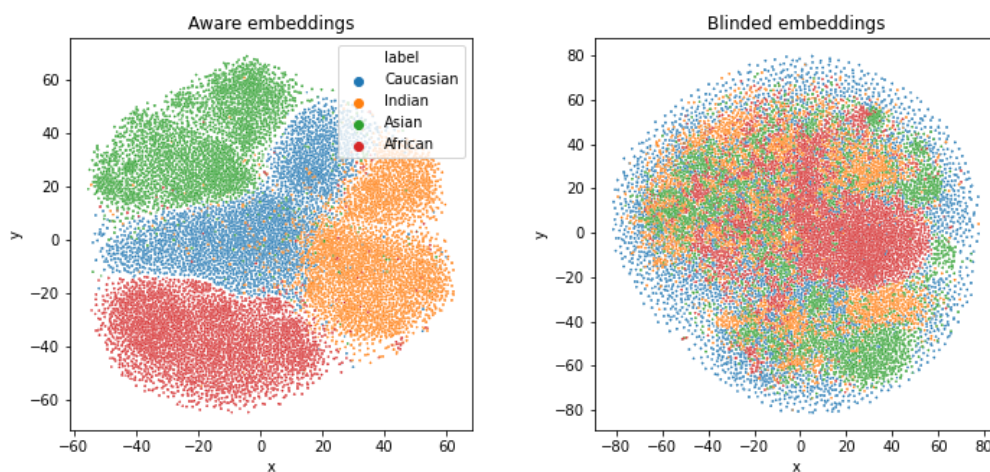


Figure 2: tSNE plots of the two embeddings

Cluster score	aware	blinded
Silhouette score	0.063	-0.014
Calinski-Harabasz score	1814	0
Davies-Bouldin score	3.8	2.8×10^6

Table 2: Cluster scores for *aware* and *blinded* emdeddings.

Groups	aware	blinded
Total	86%	88%
Caucasian	91%	92%
Indian	86%	88%
Asian	84%	86%
African	84%	86%

Table 3: Face recognition rates of the RFW dataset for *aware* and *blinded* emdeddings. The threshold was optimized for both embeddings with respect to the total dataset.

4 Face recognition rates and bias

Face recognition rates and bias are evaluated with the RFW data set. The RFW dataset provides image (i.e. embedding) pairs corresponding to the same or to different persons. The resulting task is a binary classification of the pairs into “same” and “different”. The recognition rate is the accuracy of the corresponding classification. The feature used for the classification is the pair distance in the embedding space. Here we use the cosine distance:

$$d_{ij} = 1 - \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|} \quad (4)$$

The face recognition rates calculated in this way are shown in Tab. 3. Surprisingly, the performance increases for the blinded embeddings by about 2% for all clusters. However, the bias remains. Fig. 3 gives further insights by showing the distribution of the cosine distances for “same” and “different” pairs. The Caucasians are special in that their distributions are not significantly altered. Note that the blinding procedure leads to a better alignment of the thresholds.

5 Discussion

Most of the findings with regard to the blinding procedures are not surprising. The proposed procedure removes linear separability with the effect that linear classifiers can’t distinguish between ethnic clusters after blinding in line with the cluster scores. The fact that the ethnic clusters are well separated in the first place shows that the considered

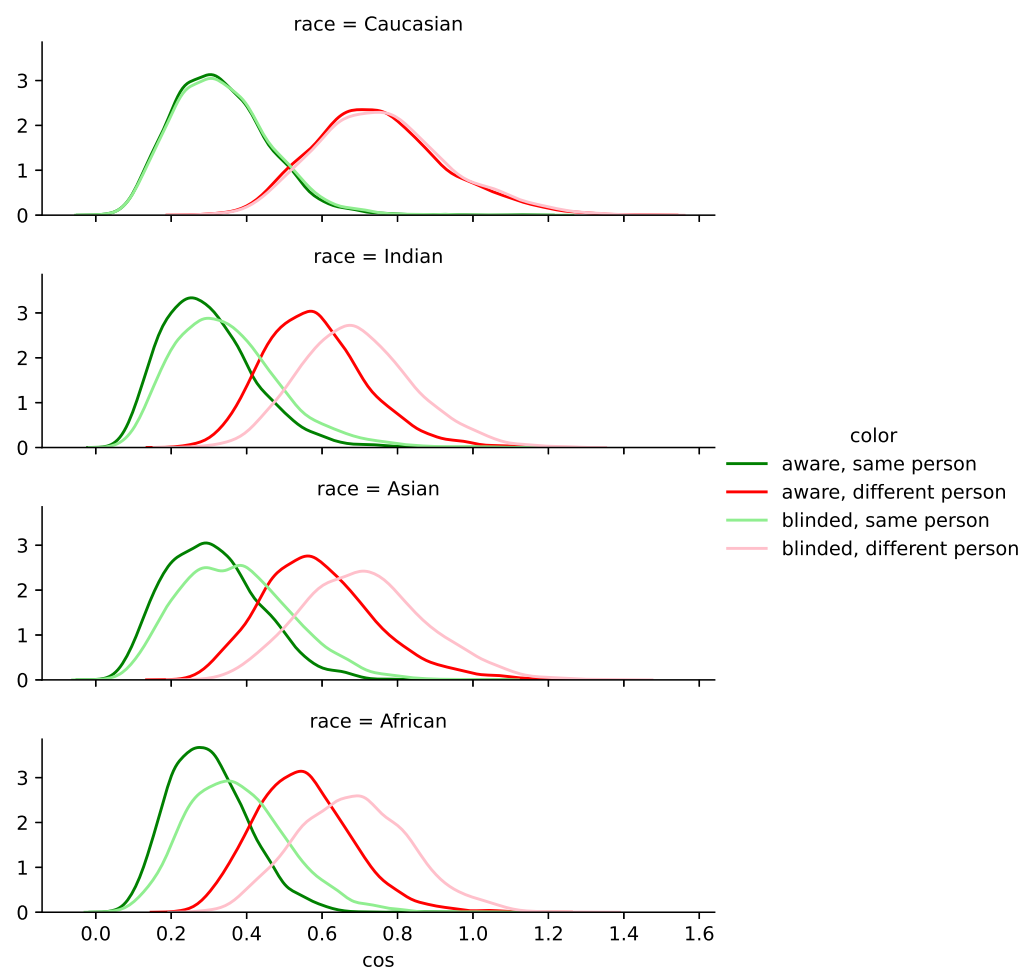


Figure 3: Distribution of cosine distances for pairs of same persons (green) and different persons (red).

model clearly distinguishes the ethnic groups. The blinding removes some of this information. The surprise comes at the end. The removal of this information actually improves overall performance - but does not remove the bias. The performance improve through this simple linear approach is very surprising. A reason might be that there are a substantial differences between the data used for training of the model and the RFW data used for present benchmark. At this stage it is not clear, whether this performance improve is accidental or generalizes to other models and other discriminatory dimensions. The Senet256 model shows a similar behavior. The effect is much smaller for the senet50ft model.