Deeply learned face representations are sparse, selective, and robust

Yi Sun[1] Xiaogang Wang[2,3] Xiaoou Tang[1,3]

[1]Department of Information Engineering, The Chinese University of Hong Kong

[2]Department of Electronic Engineering, The Chinese University of Hong Kong

[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

sy011@ie.cuhk.edu.hk xgwang@ee.cuhk.edu.hk xtang@ie.cuhk.edu.hk

Abstract

This paper designs a high-performance deep convolutional network (DeepID2+) for face recognition. It is learned with the identification-verification supervisory signal. By increasing the dimension of hidden representations and adding supervision to early convolutional layers, DeepID2+ achieves new state-of-the-art on LFW and YouTube Faces benchmarks.

Through empirical studies, we have discovered three properties of its deep neural activations critical for the high performance: sparsity, selectiveness and robustness. (1) It is observed that neural activations are moderately sparse. Moderate sparsity maximizes the discriminative power of the deep net as well as the distance between images. It is surprising that DeepID2+ still can achieve high recognition accuracy even after the neural responses are binarized. (2) Its neurons in higher layers are highly selective to identities and identity-related attributes. We can identify different subsets of neurons which are either constantly excited or inhibited when different identities or attributes are present. Although DeepID2+ is not taught to distinguish attributes during training, it has implicitly learned such high-level concepts. (3) It is much more robust to occlusions, although occlusion patterns are not included in the training set.

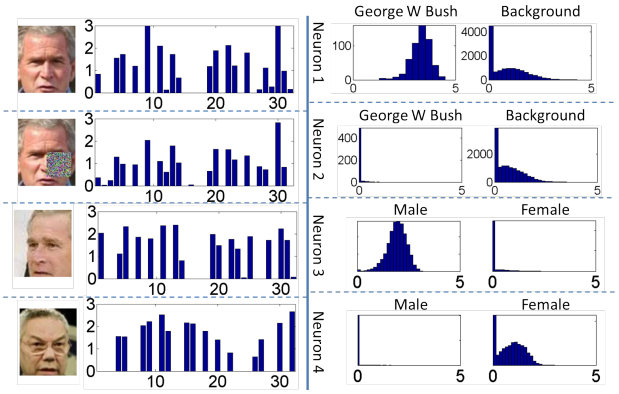


Figure 1: Left: neural responses of DeepID2+ on images of Bush and Powell. The second face is partially occluded. There are 512 neurons in the top hidden layer of DeepID2+. We subsample 32 for illustration. Right: a few neurons are selected to show their activation histograms over all the LFW face images (as background), all the images belonging to Bush, all the images with attribute Male, and all the images with attribute Female. A neuron is generally activated on about half of the face images. But it may constantly have activations (or no activation) for all the images belonging to a particular person or attribute. In this sense, neurons are sparse, and selective to identities and attributes.

1. Introduction

Face recognition achieved great progress thanks to extensive research effort devoted to this area [31, 33, 6, 24, 37, 27, 25, 23, 38]. While pursuing higher performance is a central topic, understanding the mechanisms behind it is equally important. When deep neural networks begin to approach human on challenging face benchmarks [27, 25, 23] such as LFW [13], people are eager to know what has been learned by these neurons and how such high performance is achieved. In cognitive science, there are a lot of studies [30] on analyzing the mechanisms of face processing of neurons in visual cortex. Inspired by those works, we analyze the behaviours of neurons in artificial neural networks in a attempt to explain face recognition process in deep nets, what information is encoded in neurons, and how robust they are to corruptions.

Our study is based on a high-performance deep convolutional neural network (deep ConvNet [15, 16]), referred to as DeepID2+, proposed in this paper. It is improved upon the state-of-the-art DeepID2 net [23] by increasing the dimension of hidden representations and adding supervision to early convolutional layers. The best single DeepID2+ net (taking both the original and horizontally flipped face images as input) achieves 98.70% verification accuracy on

LFW (vs. 96.72% by DeepID2). Combining 25 DeepID2+ nets sets new state-of-the-art on multiple benchmarks: 99.47% on LFW for face verification (vs. 99.15% by DeepID2 [23]), 95.0% and 80.7% on LFW for closed- and open-set face identification, respectively (vs. 82.5% and 61.9% by Web-Scale Training (WST) [28]), and 93.2% on YouTubeFaces [32] for face verification (vs. 91.4% by DeepFace [27]).

With the state-of-the-art deep ConvNets and through extensive empirical evaluation, we investigate three properties of neural activations crucial for the high performance: sparsity, selectiveness, and robustness. They are naturally owned by DeepID2+ after large scale training on face data, and we did NOT enforce any extra regularization to the model and training process to achieve them. Therefore, these results are valuable for understanding the intrinsic properties of deep networks.

It is observed that the neural activations of DeepID2+ are moderately sparse. As examples shown in Fig. 1, for an input face image, around half of the neurons in the top hidden layer are activated. On the other hand, each neuron is activated on roughly half of the face images. Such sparsity distributions can maximize the discriminative power of the deep net as well as the distance between images. Different identities have different subsets of neurons activated. Two images of the same identity have similar activation patterns. This motivates us to binarize the neural responses in the top hidden layer and use the binary code for recognition. Its result is surprisingly good. Its verification accuracy on LFW only slightly drops by 1% or less. It has significant impact on large-scale face search since huge storage and computation time is saved. This also implies that binary activation patterns are more important than activation magnitudes in deep neural networks.

Related to sparseness, it is also observed that neurons in higher layers are highly selective to identities and identity-related attributes. When an identity (who can be outside the training data) or attribute is presented, we can identify a subset of neurons which are constantly excited and also can find another subset of neurons which are constantly inhibited. A neuron from any of these two subsets has strong indication on the existence/non-existence of this identity or attribute, and our experiment show that the single neuron alone has high recognition accuracy for a particular identity or attribute. In other words, neural activations have sparsity on identities and attributes, as examples shown in Fig. 1. Although DeepID2+ is not taught to distinguish attributes during training, it has implicitly learned such high-level concepts. Directly employing the face representation learned by DeepID2+ leads to much higher classification accuracy on identity-related attributes than widely used handcrafted features such as high-dimensional LBP [6].

Our empirical study shows that neurons in higher layers are much more robust to image corruption in face recognition than handcrafted features such as high-dimensional LBP or neurons in lower layers. As an example shown in Fig. 1, when a face image is partially occluded, its binary activation patterns remain stable, although the magnitudes could change. We conjecture the reason might be that neurons in higher layers capture global features and are less sensitive to local variations. DeepID2+ is trained by natural web face images and no artificial occlusion patterns were added to the training set.

2. Related work

Only very recently, deep learning achieved great success on face recognition [27, 25, 23] and significantly outperformed systems using low level features [2, 22, 6, 4]. There are two notable breakthroughs. The first is large-scale face identification with deep neural networks [27, 25]. By classifying face images into thousands or even millions of identities, the last hidden layer forms features highly discriminative to identities. The second is supervising deep neural networks with both face identification and verification tasks [23]. The verification task minimizes the distance between features of the same identity, and decreases intra-personal variations [23]. By combining features learned from a variety of selected face regions, [23] achieved the previous state-of-the-art (99.15%) face verification on LFW.

Attribute learning is an active topic [9, 21, 20, 36]. There have been works on first learning attribute classifiers and using attribute predictions for face recognition [14, 7]. What we have tried in this paper is the inverse, by first predicting the identities, and then using the learned identity-related features to predict attributes.

Sparse representation-based classification [33, 34, 35, 8] was extensively studied for face recognition with occlusions. Tang *et al*. [29] proposed Robust Boltzmann Machine to distinguish corrupted pixels and learn latent representations. These methods designed components explicitly handling occlusions, while we show that features learned by DeepID2+ have implicitly encoded invariance to occlusions. This is naturally achieved without adding regulation to models or artificial occlusion patterns to training data.

Training deep neural networks with layer-wise supervisory signals was proposed and analysed by Lee *et al*. [17]. Our DeepID2+ nets are supervised in a similar way, but with different net structures and supervisory signals. The binarization of deep ConvNet features has been found to keep performance in object recognition [1], while we focus on face recognition and find the more interesting moderate sparsity property.

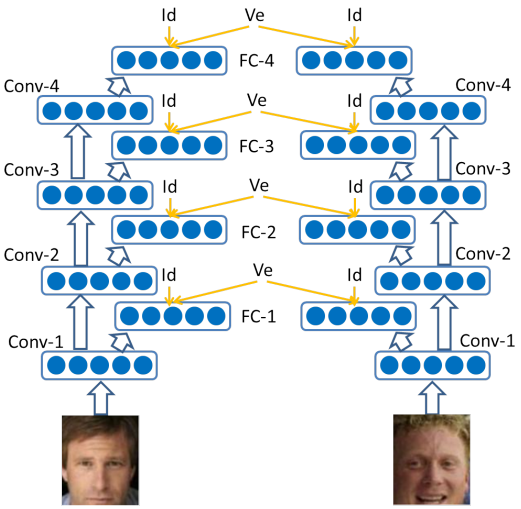


Figure 2: DeepID2+ net and supervisory signals. Conv-n deneotes the n-th convolutional layer (with max-pooling). FC-n denotes the n-th fully connected layer. Id and Ve denote the identification and verification supervisory signals. Blue arrows denote forward-propagation. Yellow arrows denote supervisory signals. Nets in the left and right are the same DeepID2+ net with different input faces.

3. DeepID2+ nets

DeepID2+ nets are inherited from DeepID2 [23] with four convolutional layers, the first three of which are followed by max-pooling, and 55×47 and 47×47 input dimensions for rectangle and square face regions, respectively. However, DeepID2+ nets make three improvements over DeepID2 as following. First, DeepID2+ nets are larger with 128 feature maps in each of the four convolutional layers. The final feature representation is also increased to 512 dimensions. Second, our training data is enlarged by merging the CelebFaces+ dataset[25], the WDRef dataset [5], and some newly collected identities exclusive from LFW. The larger DeepID2+ net is trained with around $290,000$ face images from $12,000$ identities compared to $160,000$ images from $8,000$ identities used to train the DeepID2 net. Third, we enhance the supervision by connecting a 512-dimensional fully-connected layer to each of the four convolutional layers (after max-pooling for the first three convolutional layers), denoted as FC-n for $n = 1, 2, 3, 4$, and supervise these four fully-connected layers with the identification-verification supervisory signals [23] simultaneously as shown in Fig. 2.

4. High-performance of DeepID2+ nets

To verify the improvements, we train 25 DeepID2+ nets taking the same 25 face regions selected by DeepID2 [23] and test on the LFW face verification task [13]. Features in the FC-4 layer of DeepID2+ are extracted, based on which

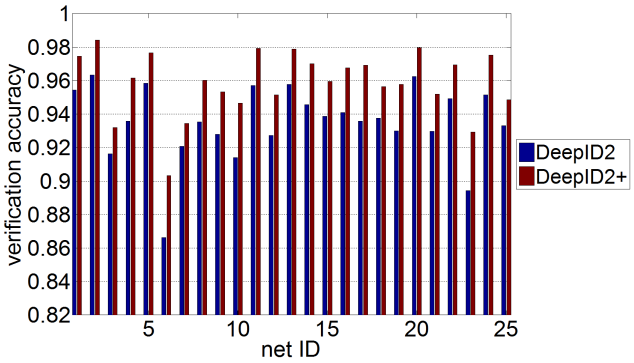


Figure 3: Comparison of face verification accuracies on LFW with ConvNets trained on 25 face regions given in DeepID2 [23]

Joint Bayesian [5] is trained on 2000 people in our training set (exclusive from people in LFW) for face verification. The comparison between the 25 deep ConvNets is shown in Fig. 3 (features are extracted on either the original or the horizontally flipped face regions as shown in Fig. 2 in [23] in the comparison). DeepID2+ nets improve approximately 2% face verification accuracy on average over DeepID2.

When combining FC-4 layer features extracted from all the 25 face regions and their horizontally flipped counterparts with the 25 DeepID2+ nets, respectively, we achieve $\mathbf{99.47}\%$ and $\mathbf{93.2}\%$ face verification accuracies on LFW and YouTube Faces datasets, respectively. Tab. 1 and Tab. 2 are accuracy comparisons with the previous best results on the two datasets. Fig. 4 and Fig. 5 are the ROC comparisons. Our DeepID2+ nets outperform all the previous results on both datasets. There are a few wrongly labeled test face pairs in LFW and YouTubeFaces. After correction, our face verification accuracy increases to 99.52% on LFW and 93.8% on YouTubeFaces.

Face identification is a more challenging task to evaluate high-performance face recognition systems [28]. Therefore we further evaluate the 25 DeepID2+ nets on the closed- and open-set face identification tasks on LFW, following the protocol in [3]. The closed-set identification reports the Rank-1 identification accuracy while the open-set identification reports the Rank-1 Detection and Identification rate (DIR) at a 1% False Alarm Rate (FAR). The comparison results are shown in Tab. 3. Our results significantly outperform the previous best [28] with $\mathbf{95.0}\%$ and $\mathbf{80.7}\%$ closed and open-set identification accuracies, respectively.

5. Moderate sparsity of neural activations

Neural activations are moderately sparse in both the sense that for each image, there are approximately half of the neurons which are activated (with positive activation values) on it, and for each neuron, there are approximately half of the images on which it is activated. The moderate

Table 1: Face verification on LFW.

method	accuracy (%)
High-dim LBP [6]	95.17 ± 1.13
TL Joint Bayesian [4]	96.33 ± 1.08
DeepFace [27]	97.35 ± 0.25
DeepID [25]	97.45 ± 0.26
GaussianFace [19]	98.52 ± 0.66
DeepID2 [23]	99.15 ± 0.13
DeepID2+	$\mathbf{99.47 \pm 0.12}$

Table 2: Face verification on YouTube Faces.

method	accuracy (%)
LM3L [12]	81.3 ± 1.2
DDML (LBP) [11]	81.3 ± 1.6
DDML (combined) [11]	82.3 ± 1.5
EigenPEP [18]	84.8 ± 1.4
DeepFace-single [27]	91.4 ± 1.1
DeepID2+	$\mathbf{93.2 \pm 0.2}$

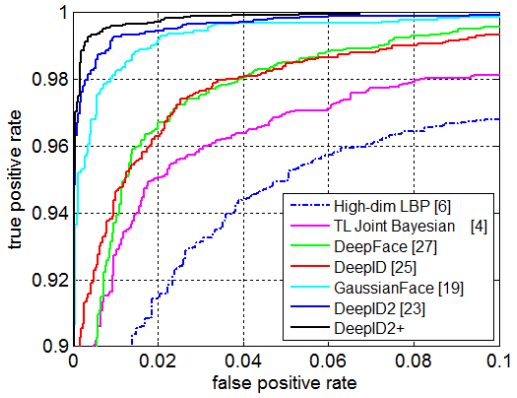


Figure 4: ROC of face verification on LFW. Best viewed in color.

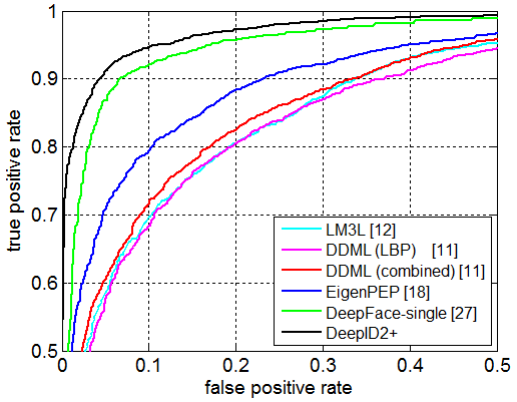


Figure 5: ROC of face verification on YouTube Faces. Best viewed in color.

sparsity on images makes faces of different identities maximally distinguishable, while the moderate sparsity

Table 3: Closed- and open-set identification tasks on LFW.

method	Rank-1 (%)	DIR @ 1% FAR (%)
COTS-s1 [3]	56.7	25
COTS-s1+s4 [3]	66.5	35
DeepFace [27]	64.9	44.5
WST Fusion [28]	82.5	61.9
DeepID2+	**95.0**	**80.7**

on neurons makes them to have maximum discrimination abilities. We verify this by calculating the histogram of the activated neural numbers on each of the $46,594$ images in our validating dataset (Fig. 6 left), and the histogram of the number of images on which each neuron are activated (Fig. 6 right). The evaluation is based on the FC-4 layer neurons in a single DeepID2+ net taking the entire face region as input. Compared to all 512 neurons in the FC-4 layer, the mean and standard deviation of the number of activated neurons on images is 292 ± 34, while compared to all $46,594$ validating images, the mean and standard deviation of the number of images on which each neuron are activated is $26,565 \pm 5754$, both of which are approximated centered at half of all neurons/images. Our experiments also show that the sparsity level is not affected by the dropout rate [10]. We take dropout of FC-n layer neurons during training. The moderate sparsity property holds for different dropout rates as well as without dropout learning. 50% dropout rate is chosen for DeepID2+.

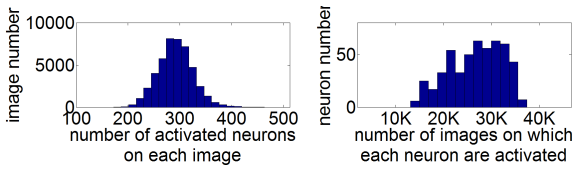


Figure 6: Left: the histogram of the number of activated neurons for each of the validating images. Right: the histogram of the number of images on which each neuron is activated.

We further verify that the activation patterns, *i.e.*, whether neurons are activated, are more important than precise activation values. We convert neural activations to binary code by thresholding and compare its face verification ability on LFW to that of the original representation. As shown in Tab. 4, the binary representation, when coupled with Joint Bayesian, sacrifices 1% or less accuracies (97.67% and 99.12% with a single net or combining 25 nets, respectively). More interestingly, the binary code can still achieve 96.45% and 97.47% accuracy with a single net or combining 25 nets, respectively, even by directly calculating the Hamming distances. This shows that the state of excitation or inhibition of neurons already contains

Table 4: Comparison of the original DeepID2+ features and its binarized representation for face verification on LFW. The first two rows of results are accuracies of the original (real-valued) FC-4 layer representation of a single net (real single) and of the 25 nets (real comb.), respectively, with Joint Bayesian as the similarity metrics. The last two rows of results are accuracies of the corresponding binary representations, with Joint Bayesian or Hamming distance as the similarity metrics, respectively.

	Joint Bayesian (%)	Hamming distance (%)
real single	98.70	N/A
real comb.	99.47	N/A
binary single	97.67	96.45
binary comb.	99.12	97.47

the majority of discriminative information. Binary code is economic for storage and fast for image search. We believe this would be an interesting direction of future work.

6. Selectiveness on identities and attributes

6.1. Discriminative power of neurons

We test DeepID2+ features for two binary classification tasks. The first is to classify the face images of one person against those of all the other people or the background. The second is to classify a face image as having an attribute or not. DeepID2+ features are taken from the FC-4 layer of a single DeepID2+ net on the entire face region and its horizontally flipped counterpart, respectively. The experiments are conducted on LFW [13] with people unseen by the DeepID2+ net during training. LFW is randomly split into two subsets and the cross-validation accuracies are reported. The accuracies are normalized w.r.t. the image numbers in the positive and negative classes. We also compare to the high-dimensional LBP features [6] with various feature dimensions. As shown in Fig. 7, DeepID2+ features significantly outperform LBP in attribute classification (it is not surprising that DeepID2+ has good identity classification result). Fig. 8 and Fig. 9 show identity and attribute classification accuracies with only one best feature selected. Different best features are selected for different identities (attributes). With a single feature (neuron), DeepID2+ reaches approximately 97% for some identity and attribute. This is the evidence that DeepID2+ features are identity and attribute selective. Apparently LBP does not have it.

6.2. Excitatory and inhibitory neurons

We find that the discrimination to identities and facial attributes are due to neurons' excitation and inhibition patterns on certain identities or attributes. For example,

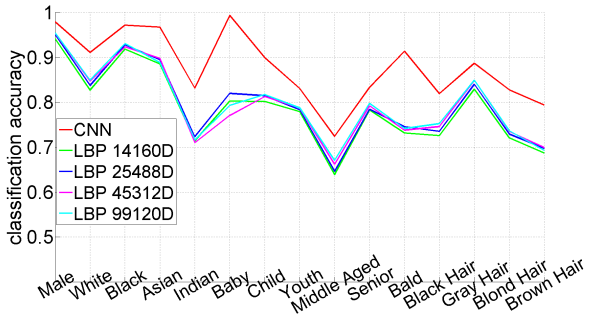


Figure 7: Accuracy comparison between DeepID2+ and LBP features for attribute classification on LFW.

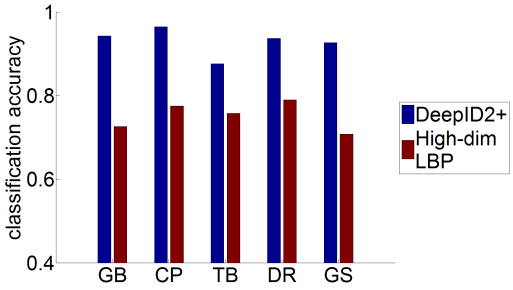


Figure 8: Identity classification accuracy on LFW with one single DeepID2+ or LBP feature. Initials of identity names are used.

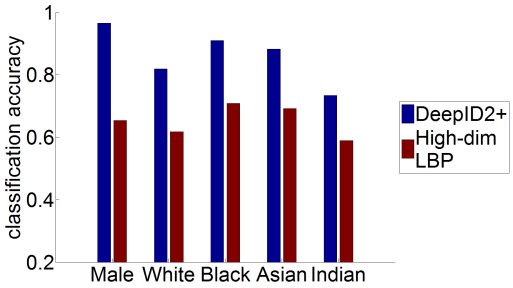


Figure 9: Attribute classification accuracy on LFW with one single DeepID2+ or LBP feature.

a neuron may be excited when it sees George Bush while becoming inhibitive when it sees Colin Powell, or a neuron may be excited for western people while being inhibitive for Asian. Fig. 10a compares the mean and standard deviation of DeepID2+ neural activations over images belonging to a particular single identity (left column) and over the remaining images (middle column), as well as showing the per-neuron classification accuracies of distinguishing each given identity from the remaining images (right column). The three identities with the most face images in LFW are evaluated (see more identities in the full version of the paper [26]). Neural orders are sorted by the mean neural activations on the evaluated identity for figures in all the three columns. For each given identity there are neurons strongly excited (*e.g.*, those with neural ID smaller than 200) or inhibited (*e.g.*, those with neural ID larger than 600).

For the excited neurons, their activations are distributed in higher values, while other images have significantly lower mean values on these neurons. Therefore, the excitatory neurons can easily distinguish an identity from others, which is verified by their high classification accuracies shown by the red dots with small neural IDs in figures in the right column.

For neurons ranked in the middle (*e.g.*, those with neural ID around 400), their activation distributions on the given identity are largely overlapped with those on other identities. They have weak discrimination abilities for the given identity, verified by the low accuracies of the red and blue dots near the junction of the two colors. The excitation or inhibition state of these neurons has much uncertainty.

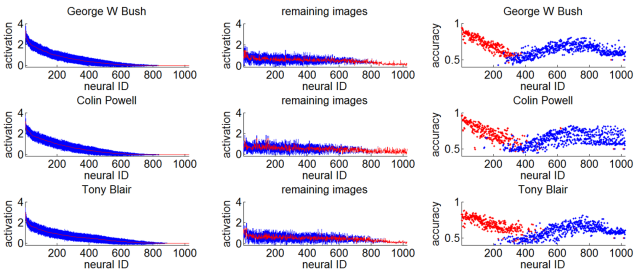
When mean activations further decrease (*e.g.*, neural ID above 600), the neurons demonstrate inhibitory properties, and are seldom activated for the given identity compared to others. These inhibitory neurons also have discrimination abilities with relatively high classification accuracies.

However, similar phenomena cannot be found on LBP features as shown in Fig. 10b. The activation distributions of LBP features on given identities and the remaining images are overlapped for all features. A LBP feature with high responses on images belonging to an identity also has high responses on other images. Compared to DeepID2+ neural activations, LBP features have much lower classification accuracies, the majority of which are accumulated on the 50% random guess line. The same conclusion can applied to attributes shown in Fig. 11a and Fig. 11b (see more examples and discussions of attributes in the full version [26]).

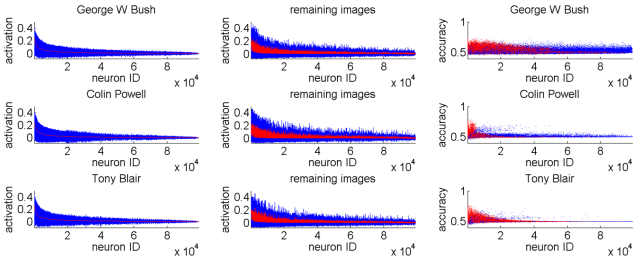
6.3. Neural activation distribution

Fig. 12 and Fig. 13 show examples of the histograms of neural activations over given identities or attributes. Fig. 12 also shows the histograms over all images of five randomly selected neurons in the first row. For each neuron, approximately half of its activations are zero (or close to zero) and the other half have larger values. In contrast, the histograms over given identities exhibit strong selectiveness. Some neurons are constantly activated for a given identity, with activation histograms distributed in positive values, as shown in the first row of histograms of each identity in Fig. 12, while some others are constantly inhibited, with activation histograms accumulated at zero or small values, as shown in the second row of histograms of each identity.

For attributes, in each column of Fig. 13a and 13b, we show histograms of a single neuron over a few attributes, *i.e.*, those related to sex and race, respectively. The neurons are selected to be excitatory (in red frames) or inhibitory (in green frames) and can best classify the attribute shown in the left of each row. As shown in these figures, neurons



(a) DeepID2+ neural activation distributions and per-neuron classification accuracies.



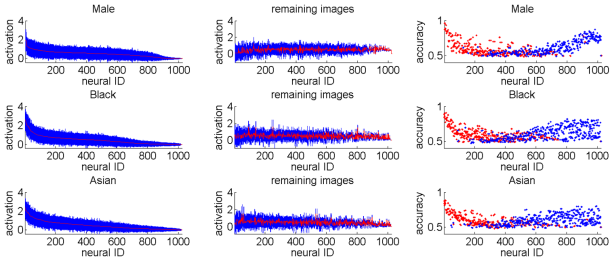
(b) LBP feature activation distributions and per-feature classification accuracies.

Figure 10: Comparison of distributions of DeepID2+ neural and LBP feature activations and per-neuron (feature) classification accuracies for the top three people with the most face images in LFW. Left column: mean and standard deviations of neural (feature) activations on images belonging to a single identity. Mean is represented by a red line while standard deviations are represented by vertical segments between (mean − standard deviation) and (mean + standard deviation). Neurons (features) are sorted by their mean activations on the given identity. Middle column: mean and standard deviations of neural (feature) activations on the remaining images. Neural (feature) orders are the same as those in the left column. Right column: per-neuron (feature) classification accuracies on the given identity. Neural (feature) orders are the same as those in the left and middle columns. Neurons (features) activated and inhibited for a given identity are marked as red and blue dots, respectively.

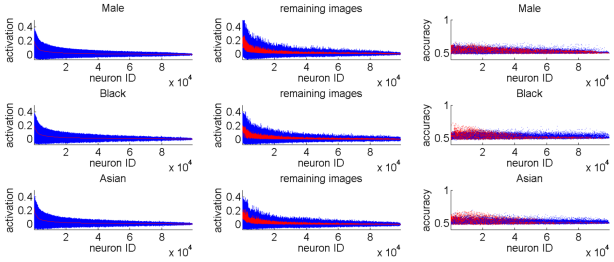
exhibit strong selectiveness (either activated or inhibited) to certain attributes, in which the neurons are activated (inhibited) for the given attribute while inhibited (activated) for the other attributes in the same category. In the full version of the paper [26] we show distribution histograms over more identities and attributes.

7. Robustness of DeepID2+ features

We test the robustness of DeepID2+ features on face images with occlusions. In the first setting, faces are partially occluded by 10% to 70% areas, as illustrated in Fig. 14 first row. In the second setting, faces are occluded by random blocks of 10×10 to 70×70 pixels in size,



(a) DeepID2+ neural activation distributions and per-neuron classification accuracies.



(b) LBP feature activation distributions and per-feature classification accuracies.

Figure 11: Comparison of distributions of DeepID2+ neural and LBP feature activations and per-neuron (feature) classification accuracies of face images of particular attributes in LFW. Figure description is the same as Fig. 10.

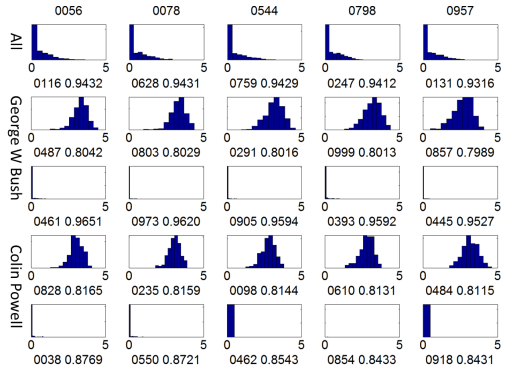
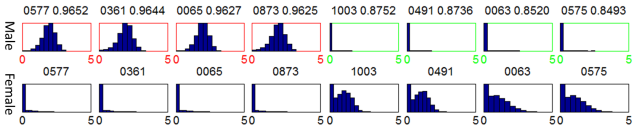
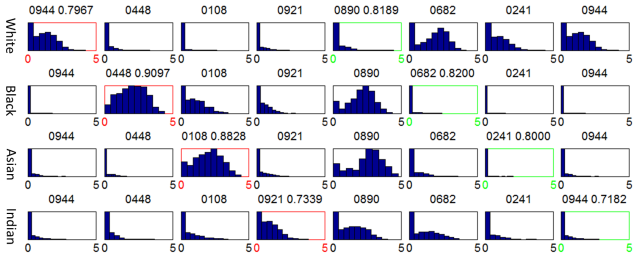


Figure 12: Histogram of neural activations. First row: activation histograms over all face images of five randomly selected neurons, with neural ID labeled above each histogram. Second to the last row: activation histograms over the two people with the most face images in LFW. For each person, histograms of five excitatory neurons (second and fourth rows) and five inhibitory neurons (third and fifth rows) with the highest binary classification accuracies of distinguishing the given identity and the remaining images are shown. People names are given in the left of every two rows. Neural ID and classification accuracies are shown above each histogram.

as illustrated in Fig. 14 second row. In the occlusion experiments, DeepID2+ nets and Joint Bayesian models are learned on the original face images in our training



(a) Histogram of neural activations over sex-related attributes (Male and Female).



(b) Histogram of neural activations over race-related attributes, i.e., White, Black, Asian, and Indian.

Figure 13: Histogram of neural activations over attributes. Each column of Fig. 13a and Fig. 13b shows histograms of a single neuron over each of the attributes given in the left, respectively. Histograms of excitatory and inhibitory neurons which best distinguish each attribute from the remaining images are shown, and are framed in red and green, respectively, with neural ID and classification accuracies shown above each histogram. The other histograms are framed in black with only neural ID above.

set without any artificially added occlusions, while the occluded faces are only used for test. We also test the high-dimensional LBP features plusing Joint Bayesian models [6] for comparison. Fig. 15 compares the face verification accuracies of DeepID2+ and LBP features on LFW test set [13] with varying degrees of partial occlusion. The DeepID2+ features are taken from the FC-1 to FC-4 layers with increasing depth in a single DeepID2+ net taking the entire face region as input. We also evaluate our entire face recognition system with 25 DeepID2+ nets. The high-dimensional LBP features compared are $99,120$ dimensions extracted from 21 facial landmarks. As shown in Fig. 15, the performance of LBP drops dramatically, even with slight 10% - 20% occlusions. In contrast, for the DeepID2+ features with two convolutions and above (FC-2, FC-3, and FC-4), the performance degrades slowly in a large range. Face verification accuracies of DeepID2+ are still above 90% when 40% of the faces are occluded (except FC-1 layer), while the performance of LBP features has dropped below 70%. The performance of DeepID2+ only degrades quickly with over 50% occlusions, when the critical eye regions are occluded. It also shows that features in higher layers (which are supposed to be more globally distributed) are more robust to occlusions, while both LBP and FC-1 are local features, sensitive to occlusions. Combining DeepID2+ nets extracted from 25 face regions achieves the most robustness with 93.9% face verification accuracy with

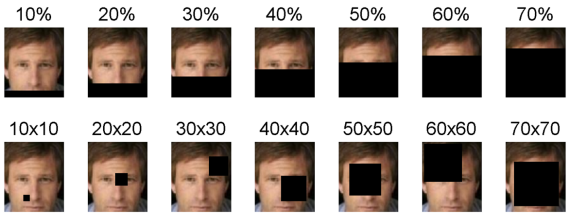


Figure 14: The occluded images tested in our experiments. First row: faces with 10% to 70% areas occluded, respectively. Second row: faces with 10×10 to 70×70 random block occlusions, respectively.

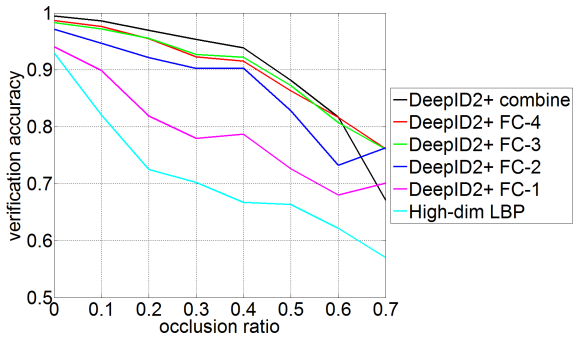


Figure 15: Face verification accuracies of DeepID2+ and high-dimensional LBP on LFW with partial occlusions. The red, green, blue, and magenta curves evaluate the features of a single DeepID2+ net, extracted from various network depth (from FC-4 to FC-1 layer). We also evaluate the combination of 25 DeepID2+ net FC-4 layer features, shown by the black curve.

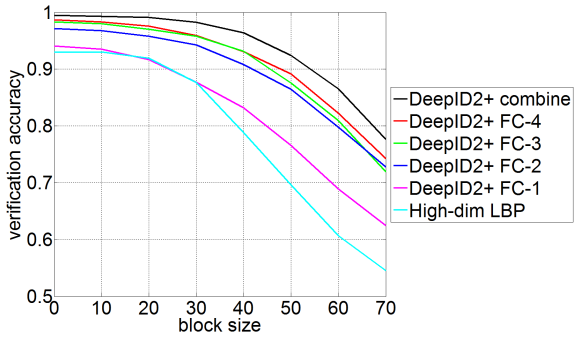


Figure 16: Face verification accuracies of DeepID2+ and high-dimensional LBP on LFW with random block occlusions. Curve description is the same as Fig. 15.

40% occlusions and 88.2% accuracy even only showing the forehead and hairs.

We also evaluate face verification of DeepID2+ and LBP features over face images with random block occlusions, with $n \times n$ block size for $n = 10$ to 70, respectively. This setting is challenging since the positions of the occluded regions in two faces to be verified are generally different. Therefore images of the same person would look much different in the sense of pixel distances. Fig. 16 shows

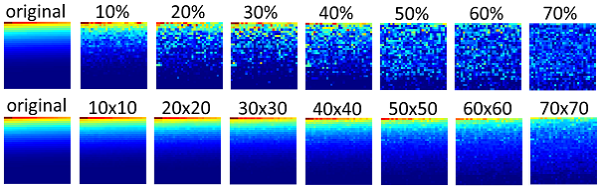


Figure 17: Mean neural activations over images of George Bush with partial (first row) or random block (second row) occlusions as illustrated in Fig. 14. Neurons are sorted by their mean activations on the original images of Bush. Activation values are mapped to a color map with warm colors indicating positive values and cool colors indicating zero or small values.

the comparison results, the accuracies of LBP features begin to drop quickly when block sizes are greater than 20×20, while DeepID2+ features (except FC-1) maintain the performance in a large range. With 50×50 block occlusions, the performance of LBP features has dropped to approximately 70%, while the FC-4 layer of a single DeepID2+ net still has 89.2% accuracy, and the combination of 25 DeepID2+ nets has an even higher 92.4% accuracy. Again, the behavior of features in the shallow FC-1 layer are closer to LBP features. The above experiments show that it is the deep structure that makes the neurons more robust to image corruptions. Such robustness is inherent in deep ConvNets without explicit modelings.

Fig. 17 shows the mean activations of FC-4 layer neurons over images of a single identity (George Bush) with various degrees of partial and random block occlusions, respectively. The neurons are ordered according to their mean activations on the original images. For both types of occlusions, activation patterns keep largely unchanged until a large degree of occlusions. See examples of more identities in the full version [26].

8. Conclusion

This paper designs a high-performance DeepID2+ net which sets new sate-of-the-art on LFW and YouTube Faces for both face identification and verification. Through empirical studies, it is found that the face representations learned by DeepID2+ are moderately sparse, highly selective to identities and attributes, and robust to image corruption. In the past, many research works have been done aiming to achieve such attractive properties by explicitly adding components or regularizations to their models or systems. However, they can be naturally achieved by the deep model through large-scale training. This work not only significantly advances the face recognition performance, but also provides valuable insight to help people to understand deep learning and its connection with many existing computer vision researches such as sparse representation, attribute learning and occlusion handling.

References

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *Proc. ECCV*, 2014.

[2] T. Berg and P. Belhumeur. Tom-vs-Pete classifiers and identity-preserving alignment for face verification. In *Proc. BMVC*, 2012.

[3] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *TR MSU-CSE-14-1*, 2014.

[4] X. Cao, D. Wipf, F. Wen, and G. Duan. A practical transfer learning algorithm for face verification. In *Proc. ICCV*, 2013.

[5] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. ECCV*, 2012.

[6] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013.

[7] J. Chung, D. Lee, Y. Seo, and C. D. Yoo. Deep attribute networks. In *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2012.

[8] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *Proc. CVPR*, 2011.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. CVPR*, 2009.

[10] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[11] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Proc. CVPR*, 2014.

[12] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Proc. ACCV*, 2014.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. ICCV*, 2009.

[15] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1:541–551, 1989.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[17] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. Technical report, arXiv:1409.5185, 2014.

[18] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-pep for video face recognition. 2014.

[19] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. In *Proc. AAAI*, 2015.

[20] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *Proc. ICCV*, 2013.

[21] D. Parikh and K. Grauman. Relative attributes. In *Proc. ICCV*, 2011.

[22] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *Proc. BMVC*, 2013.

[23] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. NIPS*, 2014.

[24] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Proc. ICCV*, 2013.

[25] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014.

[26] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. Technical report, arXiv:1412.1265, 2014.

[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014.

[28] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. Technical report, arXiv:1406.5266, 2014.

[29] Y. Tang, R. Salakhutdinov, and G. Hinton. Robust boltzmann machines for recognition and denoising. In *Proc. CVPR*, 2012.

[30] D. Y. Tsao and M. S. Livingstone. Neural mechanisms for face perception. *Annu Rev Neurosci*, 31:411–438, 2008.

[31] X. Wang and X. Tang. A unified framework for subspace face recognition. *PAMI*, 26:1222–1228, 2004.

[32] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. CVPR*, 2011.

[33] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31:210–227, 2009.

[34] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *Proc. ECCV*, 2010.

[35] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In *Proc. ICCV*, 2011.

[36] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose aligned networks for deep attribute modeling. In *Proc. CVPR*, 2014.

[37] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *Proc. ICCV*, 2013.

[38] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning and disentangling face representation by multi-view perceptron. In *Proc. NIPS*, 2014.