Visualizing and Quantifying Discriminative Features for Face Recognition

Gregory Castañón and Jeffrey Byrne
Systems & Technology Research, Woburn MA

Abstract— **Deep convolutional networks have generated significant performance improvements in the domain of face recognition. However, these improvements do not provide insight into which facial features lead to classification decisions. In this paper, we explore the problem of visualizing discriminative information in faces, to show which properties of images and subjects influence classification. We compare six different techniques for computing a network saliency map, which identifies influential local features in an image, using a metric called the "hiding game" to directly evaluate these techniques on classification performance. Results show that contrastive excitation backprop (cEBP) [26] best localizes features that lead to face identification. However, these maps are nearly identical across subjects, which can result in an unstable network saliency map. We introduce a robust improvement called truncated cEBP and demonstrate the capability to predict the performance of a given map. Our evaluation provides the first application of network saliency to face recognition, and we provide a robust new tool for face recognition analysts to explore which facial regions lead to changes in match scores.**

I. Introduction

Convolutional networks have become the standard representation for face recognition. While representations demonstrate powerful performance, the networks are opaque and do not provide insight into which facial features lead to classification decisions. Recently, an increased emphasis has been put on network transparency in order to better understanding factors within images, subject classes and datasets that affect network performance. *Network Attention* is a specialization of deep network visualization [24][23][17] that reveals which areas of an image influence the classification decision.

In this paper, we evaluate six different approaches for estimating network attention in the domain of face recognition. We focus on the properties of the excitory attention signal, which captures regions of an image that are most predictive of a specific class. We ask the following questions: (i) Can the importance of the highlighted areas of the image be confirmed by classification performance? (ii) Can the signal be used to discriminate between difference classes? (iii) How do different approaches to network attention map generation compare to predict classification performance? To investigate these questions, we leverage existing work in network attention [15][18][27][26][3][5][27]. Our primary investigation focuses on recent work on *excitation backprop* [26], which captures network saliency in a probabilistic winner-take-all framework.

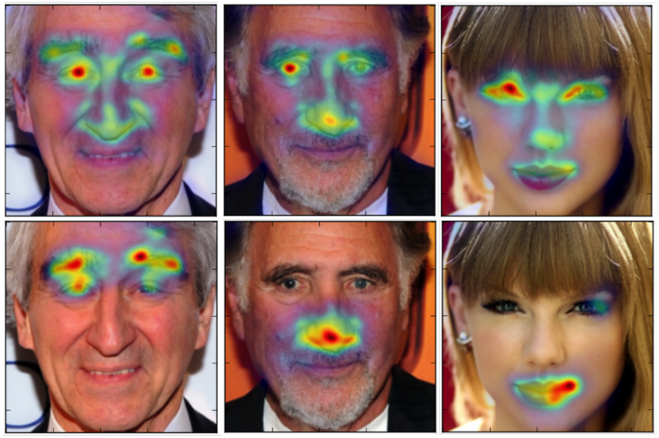

Fig. 1. We explore visualizing and quantifying discriminative features for face recognition. (top row) A network attention map localizes excitory features that lead to classification of a subject, (bottom row) A contrastive network attention map captures properties that distinguish a particular subject from the training population. We use these attention maps to incrementally "hide" a face, which allows us to answer quantitatively "what is important for classification" and "what makes a subject unique?".

Figure 1 (top) shows an example of a network attention for three distinctive celebrities from the VGG-Face dataset [11]: Sam Waterston, Judd Hirsch and Taylor Swift. The top row shows the class-specific map of network attention generated using excitation backprop [26] in the form of an *attention map*. An attention map is positive signal that sums to one, which captures the relative importance or *excitation* of regions of the face for predicting a specific class. For example, attention map regions in red and yellow around the eyes show the strongest excitation, and attention maps in blue green and show the weakest excitation. This example shows that the eyes and the nose are most important for predicting the subject, and the cheeks and mouth are less important. Observe that the attention map is nearly the same for all three classes, where the eye and nose regions dominate.

Figure 1 (bottom) shows an example of a *contrastive network attention map*, generated using contrastive excitation backprop [26], which captures the excitations for a class that are not found in other classes. For example, the contrastive excitation for Taylor Swift is her lips, which are slightly

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under contract number 2014-14071600010. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

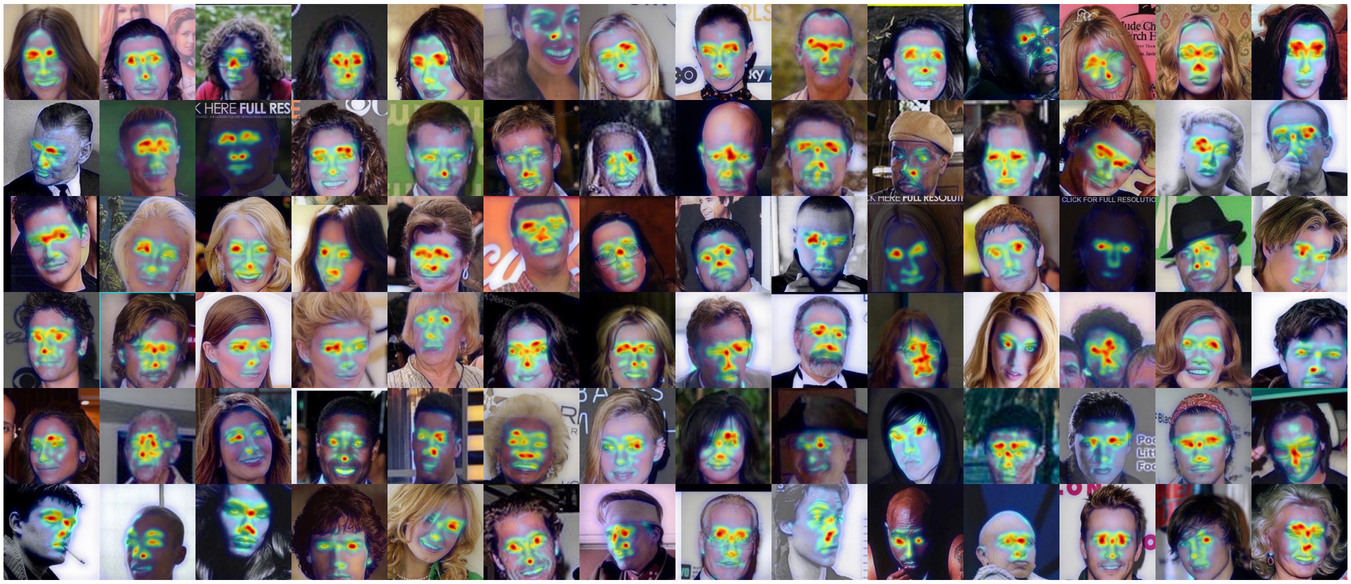

Fig. 2. Facial network attention maps computed for eighty four subjects in the VGG-Face dataset [11].

more excitory for her than other people. Similarly, Sam Waterston's eyebrows or Judd Hirsch's nose are slightly more excitory for the respective classes. The visualization of a network attention map and contrastive network attention map has the potential for use in highlighting distinctive areas the faces, providing insight into the areas of a face that a network considers exceptional or unique. In other words, "what makes you different from everyone else?".

Figure 2 shows examples of eighty four network attention maps computed using excitation backprop for subjects in the VGG-face dataset. These visualizations appear to be consistently localizing the eyes, nose and mouth, but do not provide quantitative evidence for how well the individual network attention maps are performing. There are many alternate strategies for computing network attention and all generate a visualization similar to Figure 2. How should these maps be compared and quantitatively evaluated? Are they useful for predicting classification performance?

To answer these questions, we propose an evaluation metric called the "hiding game" inspired by [5][14][13] which incrementally covers portions of an image according to the attention map, and compares network classification performance on the remaining portions of the image. This allows us to investigate the direct connection between network attention and classification performance to show which regions of the face are predictive of a class. Furthermore, this provides a quantitative evaluation to compare performance across different strategies for network attention, and provides a useful tool for analysts to quantify which facial regions lead to higher or lower match scores.

The main contributions of this paper are as follows:

- Visualizing the network attention map for a modern convolutional network trained for face recognition. As far as we know, this is the first published visualization of features for the task of face recognition.
- Visualizing the contrastive network attention map to highlight distinctive features for a specific subject.
- Proposing a new algorithm for network attention called truncated contrastive excitation backprop which addresses an observed instability for computing contrastive attention for faces.
- Comparing six different network attention maps for the task of subject identity classification in the "hiding game" and the "contrastive hiding game" to connect network attention with face recognition performance.

II. Related Work

Many approaches to estimating network attention in deep networks have been proposed. Gradient-based methods [15][18][27] attempt to compute the the derivative of the class signal with respect to the input image, while other approaches [3] modify network architectures to capture these signals. Other approaches, like excitation backprop [26] formulate the signal as the solution to a probabilistic winner-take-all problem. Feedback loops [3] for black-box methods [5] directly alter the input image to achieve maximal effect on classification performance with minimal change. These approaches have shown promise in object localization [27] in datasets like Imagenet [12], where objects are very visually distinct. Inversion methods [10] seek to recover natural images that have the same feature representation as a given image. However, the same insights have not yet been applied to fine grained categorization for face recognition.

Bottom up visual saliency has been explored in the literature [2][1][6][7]. These approaches construct a visual attention map independent of the semantic interpretation of a scene, to model the pre-attentive regions of an image important for classification. In this paper, we do not perform evaluations of a visual saliency map based on how well this predicts human performance from measurements such as eye

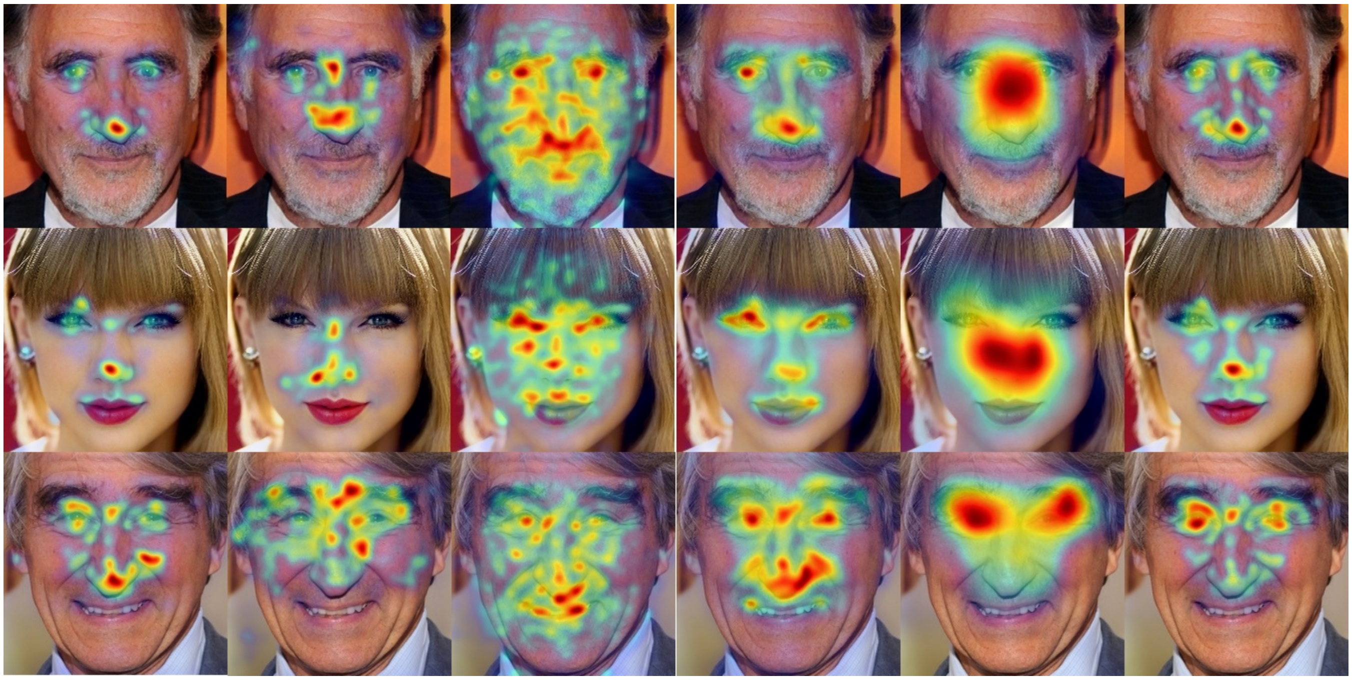

Fig. 3. Visual comparison of six different strategies for computing a network attention map. Columnwise left to right: Grad-Cam [15], Guided Grad-cam [15], Gradient-based saliency [18], Excitation Backprop (EBP) [26], Guided backprop [20] and Deconv nets [25]

tracking. Rather, we are interested in how well the saliency map predict the classification performance.

Top down visual saliency has been proposed to provide attention maps to improve classification tasks. Latent attention networks [4] treat visual saliency as a latent task, which is trained to output a mask that indicates the degree to which an input component can be replaced by noise without affecting the classification output. This is related to the hiding game, incorporating an optimal mask objective into the loss function. Fully convolutional attention networks [22] use an learned part-based attention map to improve fine grained categorization tasks. Attention transfer [16] shows that training a student network to imitate the attention maps of a teacher network can improve classification performance.

Tools for visualizing and understanding convolutional networks have been used for many tasks in computer vision. For example, these tools have been applied to texture recognition [8], fine grained categorization [9] and object classification [24]. However, this is the first paper to explore the visualization and understanding for face recognition. Prior work has considered learning to predict saliency in face images using eye tracking from human subjects [21], however this does not provide a connection with automated classification tasks from a deep network. Furthermore, prior work exists for visualizing facial attributes in the wild using deep networks [28], but not connecting to the task of face recognition.

Finally, recent work has connected visual saliency maps with metrics for classification tasks. Interpretable explanatation of black boxes [5] attempts to delete a region that will maximally affect a classification score. The authors show that the resulting mask provides the image support for classification, but this is not appropriate for fine grained categorization problems such as faces, where an ordering of the attention map is necessary. The approach in [14][13] provides a layerwise relevance propagation that provides a quantitative comparison of attention maps. This is closely related to the hiding game proposed here, however this approach replaces the most relevant areas with uniform random noise, while ours replaces the least relevant areas with the mean color in the training set for a network.

III. Network Attention

In this section, we describe the primary technique for computing the network attention map explored in this paper based on excitation backprop (EBP) [26].

We are interested in exploring the *network attention* signal in a traditional feed-forward convolutional network. This *excitory signal* $A(i,j)$ scores pixels $i, j \in \mathcal{I}$ for image $\mathcal{I}$ by their contributions to a particular task - in this case, the task of classifying $\mathcal{I}$ correctly. One approach to this problem is *excitation backprop* [26], which computes a *marginal win probability* for each neuron. Let $\mathcal{C}_i$ be the child (top-down order) neuron set of a given neuron $a_i \in \mathcal{N}$, where $\mathcal{N}$ is the set of neurons in the network. Then for each neuron $a_j \in \mathcal{C}_i$, the conditional win probability is given by:

$$P(a_j|a_i) = \begin{cases} Z_i \hat{a}_j w_{ji} & \text{if} w_{ji} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where Z_i is a normalization constant such that $\sum_j P(a_j|a_i) = 1$, $\hat{a}_j$ is the feed-forward activation of neuron a_j, and w_{ji} is the weight of the edge connecting a_j to a_i. Note that we only use $w_{ji} \geq 0$, as we are only interested in neurons that have a positive correlation with the class signal. The intuition is that the feed-forward signal $\hat{a}_j$ indicates which neurons are present/active in the image, while the weights w_{ji} indicate the relative importance of each neuron to the next layer. Beginning with the class signal at the top, this allows us to compute the excitory signal A_{ij} at the pixel level.

This process can also be used to compute *contrastive network attention* between multiple signals. We are particular interested the *contrastive signal* $C(i, j)$ which is the difference between excitory signal for a single class $s_i \in \mathcal{S}$ and the average excitory signal for all other classes $\mathcal{S} \bigcap S_i$.

$$C_i = A_i - \sum_{j \in \mathcal{S}, j \neq i} A_j \tag{2}$$

Zhang et. al [26] shows that this signal can be calculated via *contrastive excitation backpropagation* (cEBP) at any layer of the network including the input. Our goal is to use the excitory signal $A(i, j)$ and the contrastive signal $C(i, j)$ to reveal information about particular images, classes, and datasets. For additional details on the excitation backprop and contrastive excitation backprop see the discussion in [26].

This probabilistic winner-take-all process has the effect of concentrating signal, which is a desirable property for localizing important areas of an image. Once that signal is concentrated in the convolutional layers, it results in an attention map that concentrates in specific areas of the image, as opposed to being spread out like other methods such as the deconvolutional approach shown in Figure 3.

As a post-processing step, we apply a gaussian filter of $\mathcal{N}(0, k^2)$ to smooth the map, where k is equal to 2% of the image's width, or 4.5 pixels for a 224x224 image, and then normalize the mass of the signal to one. This provides a smooth map with clear peaks in the signal, but occasionally accentuates noise when the signal is low-amplitude.

IV. Network Attention Experiments

While the relative benefits of network attention maps have been discussed and visualized previously, their effect at ranking areas of an input image by impact on classification performance has not been quantized. In this section, we lay out an experimental methodology for evaluation.

A. Experimental System

As noted in Section II, there are many approaches to computing network attention maps for a given image, image class, and network. We evaluate excitation backprop [26], described above, and five other signals. Gradient-based backpropagation [18] computes the gradient w of the class signal with respect to the image, and takes the maximum value over the images' color channels. Deconvnets [24] backpropagate only the positive components of the gradient, while guided backpropagation [20] zeroes out all components either the feed-forward signal or the feedbackward signal are negative. Grad-CAM [15] extends class activation maps (CAM) [27] by spatially pooling feature maps using Global Average Pooling. The authors also present a method of localizing the Grad-CAM signal by multiplying it with the guided backpropagation signal, called Guided Grad-CAM [15].

As a basis for experimental comparison, we used the VGG-Face [11] network. This network is a VGG-16 topology with softmax loss, trained on 2.6 million images of 2,622 unique celebrities, using the training strategy originally defined in [19]. From the test set, we randomly selected 1000 images and computed the six signals for each of these images. We evaluate the excitory signal qualitatively in Section IV-B, and quantitatively by its effect on classification performance using "the hiding game" in IV-C. For the contrastive signal, in Section V-A we explore visualizations that highlight its ability to discriminate between subjects and detect dataset bias, before evaluating its failure modes and instability. We then apply a robustified contrastive attention to show classification performance on "contrastive hiding game" in section V-C.

B. Network Attention Maps

Figure 3 shows a comparison of network attention for six different strategies, such that each column is a different strategy, and each row is the same image. Observe that the area interior to the face is highlighted across methods; the network has learned to ignore areas of an image such as the background, as may be expected. Also, most approaches focus on the nose and the eyes, while chin, cheeks and mouth are less highlighted. Anecdotally, guided methods seem to produce sharper, more disjoint maps, while deconvolutional approaches do not localize as well - this matches intuition, as guided approaches filter out the most components of the signal. Class-Activation Maps provide a significantly smoother signal than the others as a result of average pooling.

These examples confirm the hypothesis that the features used for subject classification are the regions that are commonly associated with invariant face recognition. The visualizations show that regions of hair, beards, and mouth are rarely excitory for a specific class. Instead, the regions that are excitory for subject classification are the region around the eyes, the nose and above the lips. Human visual saliency studies using eye tracking evaluations of humans looking at faces have shown remarkably similar patterns of activations [21], and these results show that the convolutional network has learned representations of similar regions.

C. The Hiding Game

The visualizations in Figure 3 show qualitative evidence that the six different methods of computing network attention maps highlight different components of a given face. Implicitly, each of these maps also provides a ranking of the importance of each pixel in the image to the class's signal. We seek to answer if this ordering of pixels correlates to the relative importance of pixels for classification.

To experimentally measure this, we adapted an approach proposed in [5][14][13], which we call "The Hiding Game". The goal of the hiding game is to iteratively obscure the least important pixels in the image sorted according to an attention map, replace the hidden pixels with a deterministic

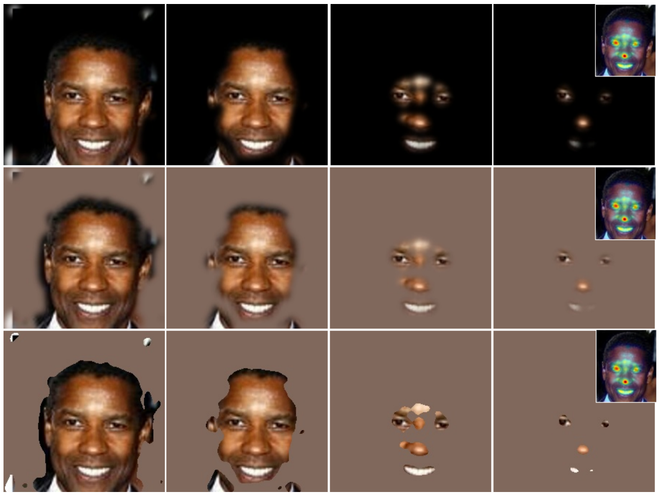

Fig. 4. The Hiding Game. A subject is iteratively hidden in order of increasing importance according to the attention map and the hidden pixels are replaced with a constant value, leaving 50%, 70%, 90% or 99% of the image hidden. (top row) Zero replacement, (middle row) Mean replacement with soft-mask or (bottom row) hard-mask.

value, then compute a classification score for the partially obscured image. For example, Figure 4 shows an example of a subject with a corresponding attention map in the upper right corner. Each column shows a different hiding threshold, which shows progressively more of the image hidden. Each row is a different strategy of replacement for the hidden pixels, replaced with zeros, replaced with the training set mean color with or without a Gaussian blur. Correct classification of a hidden image provides strong evidence that the remaining unobscured pixels are important for classification of the subject.

Figure 5 (right) shows the classification performance of these approaches. We evaluate for a test set of 1000 images collected from the VGG-face dataset, and measure the percentage of images correctly classified as a function of percentage of the image hidden using the EBP attention map. We compare the EBP attention map to a uniform random attention map. These results show that the best strategy is mean replacement with soft-mask blur (e.g. row two of Figure 4), which exhibits a large improvement over zero replacement and significant improvement over a random attention map. Furthermore, these results show how little of the image is actually important for correct classification. Using an EBP signal, we can hide 70% of the pixels in an image and still classify faces correctly over 80% of the time. This provides evidence that the attention map is capturing pixels in order of importance for subject classification.

Figure 5 (left) shows the classification performance for the six different strategies for network attention map as visualized in Figure 3. We use the same test set as Figure 5 (right), and a mean replacement soft-mask strategy for all approaches. Results show that the EBP signal provides a large improvement over the second best technique based on Guided backprop [20] and Gradient-based saliency [18]. This provides a quantitative comparison of the attention maps in Figure 3, and provides strong evidence that excitation backprop provides the best ordering of pixels for classification.

V. Contrastive Network Attention Experiments

The contrastive network attention signal C_i for the i-th subject is computed by subtracting the average attention signal for all other subjects, $A_j \ j \neq i$, from A_i. Though this signal is significantly smaller in magnitude than the network attention signal, it can highlight subtle areas that differentiate the primary subject from the population at large. In our experiments, we tried to generate stable visualizations of these areas and explore what information could be gleaned from these contrastive maps.

A. *Visualizing Contrastive Network Attention Maps*

Visualizations of contrastive network attention can provide insight into areas of an image that are uniquely important for that subject. One approach to computing contrastive network attention maps is described in section III using contrastive excitation backprop. This approach visualizes the features that are more excitory for a given subject and less excitory for other subjects in the training set. In other words, this is a visualization of the contrast between the excitory features that are present for all subjects vs. excitory features that are present for one subject.

Figure 6 shows an example of contrastive network attention. Figure 6 (a) and (d) show the smoothed attention map overlay and pixel level excitation map from EBP for the Jared Leto Class, whereas (b) and (e) show the smoothed attention map and pixel level excitation map from EBP for the not Jared Leto class. Figure 6 (c) shows the strictly positive difference between (a) and (b), such that $(c) = \texttt{max}((a) - (b),\ 0)$, forming the contrastive attention map. Figure 6 (f) shows the contrastive attention map without truncation and smoothing such that $(f) = (d) - (e)$, and (c) is (f) truncated at zero and smoothed. The map in (f) provides a slightly better visualization of the attention map.

Figure 6 shows that excitation maps for Jared Leto and not Jared Leto are nearly identical. Observe that (d) and (e) as well as (a) and (b) are indistinguishable to a human observer. This shows that to discriminate Jared Leto from anyone else requires features around the eyes, nose and mouth, whereas to distinguish that anyone is not Jared Leto requires looking at the same regions. However, these signals are not exactly identical. There are subtle differences as shown in (f)=(d)-(e), such that blue is negative, yellow is positive and green is zero. This shows that there are regions around the mouth that are slightly more excitory for Jared Leto than for anyone else, as visualized in the yellow region of Figure 6 (f). This is further visualized by truncating this signal at zero, and smoothing forming the contrastive network attention map in Figure 6 (c). The conclusion from this is that there are subtle differences that are unique to this image of Jared Leto that are not found in the rest of the training set, and this subtle difference is captured in the contrastive attention map. Examples of the contrastive network attention map are shown in Figure 1 (bottom row).

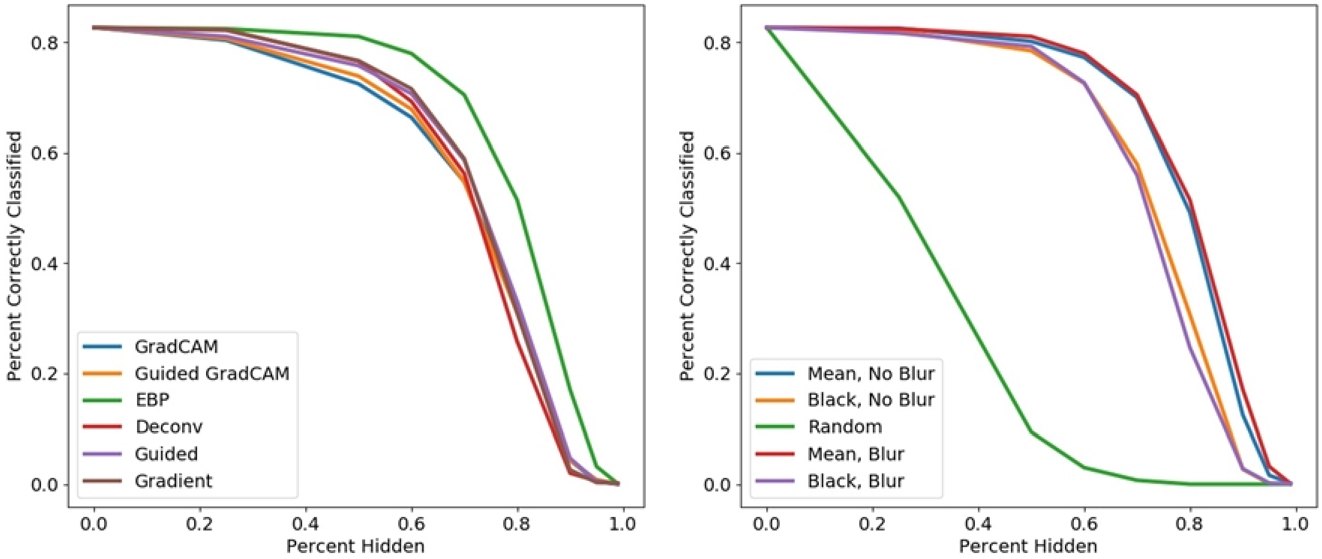

Fig. 5. Hiding game results. (left) Hiding game evaluation of six different strategies: Gradient-based saliency [18], Guided backprop [20], Deconvolutional networks [25], EBP [26], Grad-CAM [15], and Guided Grad-cam [15]. These results show EBP [26] best captures the important pixels for face identification. (right) This plot shows how performance degrades as a function of percentage of image hidden. We compare Excitation backprop with mean replacement and blurring to unblurred, zero-replacement, and a random baseline.

B. Truncated Contrastive Excitation Backprop

The visualization for contrastive network attention in section V-A focused on contrastive excitation backprop [26]. However, our experiments have shown that this technique can be noisy and unstable when applied to faces. In this section, we introduce a robust modification of this technique to address this instability called truncated cEBP.

Figure 7 shows four examples of unstable behavior of cEBP applied to faces. The second row shows the cEBP signal which exhibits contrastive attention in the hair of Taylor Swift and on the clothes on Judd Hirsch. These regions of hair and clothes have very low excitation according to the EBP signal in the bottom row. Why is the cEBP signal non-zero in these regions? Intuitively, the contrastive network attention map should capture those regions that are both *important* for classification and *unique* for a subject.

Figure 7 (third column) shows a near perfect example of Judd Hirsch. This is a near frontal image, with neutral expression and as a result there are large regions in this image with strong excitation, including portions of the temples, eyebrows and chin. Recall that the cEBP signal is constructed from the difference between the EBP for the subject and the EBP for not-Subject (Figure 6). When then not-Judd Hirsh excitory signal is subtracted, the only positive components of the difference are outside the face. This is due to the fact that the EBP signal is normalized to be a probability distribution to sum to one. When mass for the probability distribution is distributed to peripheral regions of the image such as the temples or chin, the mass must be reduced from the eyes and nose due to the normalization. However, the mass for the normalized EBP signal for the not-Judd Hirsch classifier is centered on the interior facial features only. Subtracting the excitation maps results in the interior facial features with negative excitation and the exterior facial features have positive excitation. Therefore, the cEBP signal that is strong in regions outside the face, forming an undesirable map that is "unique" but not "important" for classification.

To address this instability, we introduce a modification of the contrastive EBP called *truncated contrastive EBP*. As shown in Figure 7, the cEBP instability occurs in regions of low EBP excitation. However, these are regions are not "important" in that they do not contribute strongly to classification of the subject. Truncated cEBP computes a truncation of the EBP and negative EBP signals prior to normalization, so that low likelihood regions less than a fixed ϵ are set uniformly to zero. Formally, let P be the excitation backprop signal for a given class, and N be the negative excitation backprop signal for the class. Then, the truncated contrastive EBP (T) is:

$$T = \mathtt{max}\left(\frac{\mathtt{max}(P-\epsilon,\ 0)}{\sum \mathtt{max}(P-\epsilon,\ 0)} - \frac{\mathtt{max}(N-\epsilon,\ 0)}{\sum \mathtt{max}(N-\epsilon,\ 0)},\ 0\right) \quad (3)$$

where the function $\mathtt{max}(f-\epsilon, 0)$ truncates map f elementwise to be greater than epsilon else zero. Compare this to equation (2) for classic cEBP which does not include ϵ.

The value of epsilon is set experimentally on a validation set to be $\epsilon = 1.94e{-}5$. We compute the 70th percentile of the EBP signal, and set epsilon to be the mean of this statistic on a validation set. Intuitively, the value of epsilon corresponds to the mean EBP value that results in 70 percent of the image hidden on the validation set. This truncation forces the cEBP signal to be within the support of the excitation map.

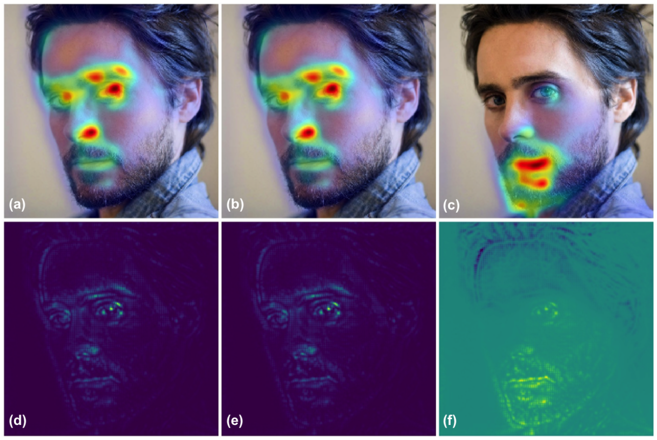

Fig. 6. Contrastive network attention map. (a) The network attention map for Jared Leto and (b) not-Jared Leto, (c) the contrastive network attention map formed from the strictly positive difference of figures (a)-(b), (d) The excitory signal for Jared Leto and (e) not-Jared Leto, (f) the contrastive excitory signal (a)-(b). The map in (c) is formed by truncating (f) at zero and smoothing.

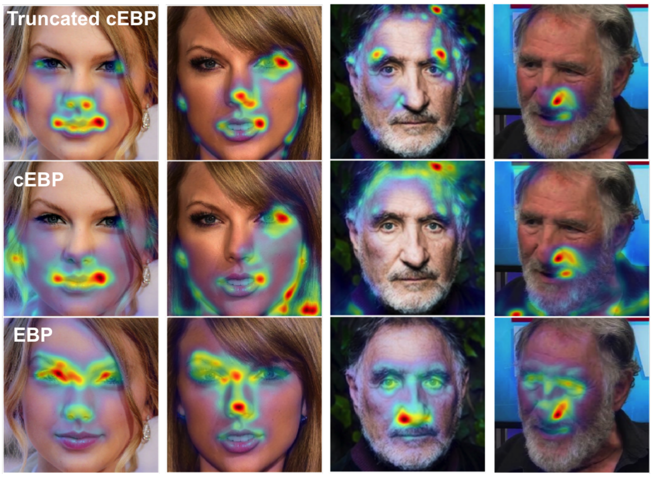

Fig. 7. Truncated contrastive excitation backprop. (top row) Truncated cEBP (middle row) cEBP (bottom row) EBP. Observe the instability in the cEBP signal in second column, second row for regions of low EBP around the hair in the bottom right.

C. The Contrastive Hiding Game

The results of the hiding game, presented in section IV-C, show that the network attention map computed by excitation backprop is the most accurate at ranking areas of an image by their importance for correct classification. However, intuitively, some images may contain information that is more *discriminative* for a particular subject than anyone else. To quantify, we propose a modification to the hiding game called the *contrastive hiding game*, which highlight the discriminative power of the contrastive signal.

In section III, we defined the contrastive network attention map for a specific class as the difference between the excitory signal for that subject and the average excitory signal for all other classes. This removes excitations common to a subject and everyone else, while leaving only differences, forming a contrastive attention map. Regions of positive difference are those facial regions that exhibit more excitation for the subject than anyone else, which suggests that these positive difference regions contain the facial features that are most unique for the subject. Furthermore, suppose we obscure a face image, such that there is more positive excitation than negative in the unobscured regions, we would expect that this image would result in a higher match score for the subject. Can we show this experimentally?

The contrastive hiding game was designed to answer this question. It is played like the hiding game, except we additionally compute the mean cEBP signal within the un-masked region using the truncated cEBP from section V-B (denoted the "cEBP Score"). We define an *activation* to be the fc-8 output of the VGG-16 topology for a subject prior to softmax layer, and *softmax score* to be the softmax normalized output for a subject. For each image, we measure the activation a_0 of the true subject without any hiding, and the activation a_m with m percent of the image hidden. We then compute the cEBP score within the unhidden region of the image. We expect images with a positive cEBP score to have lower differential activation $a_0 - a_m$, and higher absolute score a_m. In other words, when an masked image contains positive cEBP score, there are more unique excitations for the subject than any other subject, which should result in higher subject match scores. This provides a direct measurement of the effect of masking and contrastive attention on match scores.

Figure 8 shows the results of the contrastive hiding game. Figure 8 (left) shows the differential activation $(a_0 - a_m)$ for images that are 50%, 60% and 70% masked, and Figure 8 (right) shows the softmax score. For a fixed percent hidden, as the mean cEBP score increases from negative to positive, we see that mean softmax scores increase and mean differential activations decrease. This shows that cEBP is strongly correlated with match score, and that cEBP provides evidence that cEBP captures unique features that increase only the scores of the mated subject, and not others. This shows quantitatively that cEBP captures unique features that distinguish a subject from others.

VI. Conclusions

In this work, we explored the use of network attention and contrastive network attention for visualizing discriminative features for face recognition. We demonstrated through the hiding game that excitation backprop best identifies the components of an image that contribute to correct classification. Furthermore, we demonstrated with the contrastive hiding game that contrastive excitation backprop identifies those unique features that contribute to increased match scores for a given subject and network. This suggests that contrastive network attention can be used as a new tool by facial biometric analysts for adjudication of 1:1 verification and 1:N identification tasks for convolutional networks.

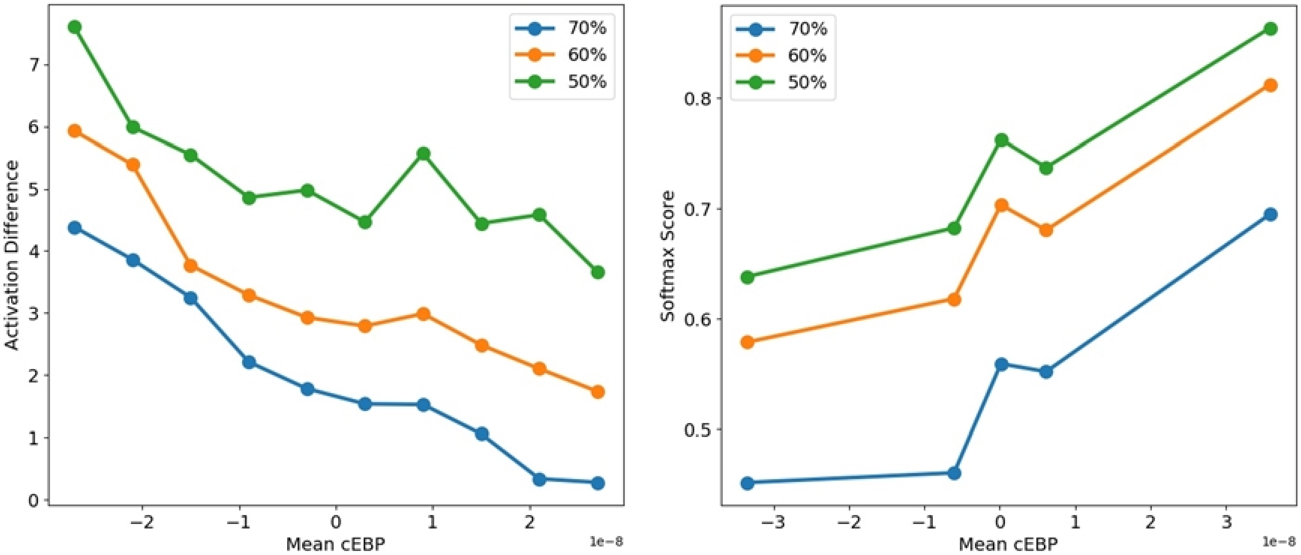

Fig. 8. Contrastive hiding game results. We compute a mean cEBP score for the test set for images that are 50, 60 and 70% obscured. (Left) The activation difference between the unhidden and hidden images as as function of mean cEBP score. (Right) The activation as a function of mean cEBP score. These results show that cEBP is strongly correlated with match score.

REFERENCES

[1] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[2] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, 2016.

[3] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, and Others. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[4] S. K. D. A. L. L. W. M. L. L. Christopher Grimm, Dilip Arumugam. Latent attention networks. In *arXiv:1706.00536v1*, 2017.

[5] R. Fong and A. Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *arXiv preprint arXiv*, 2017.

[6] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[7] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[8] T.-Y. Lin and S. Maji. Visualizing and understanding deep texture representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnns for fine-grained visual recognition. In *Transactions of Pattern Analysis and Machine Intelligence (PAMI)*.

[10] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *BMVC*, volume 1, page 6, 2015.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[13] G. M. F. K. K.-R. M. S. Bach, A. Binder and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PLOS ONE, vol. 10, no. 7,p. e0130140*, 2015.

[14] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. 08 2016.

[15] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. *1610.02391V2*, (Nips):1–5, 2016.

[16] N. K. Sergey Zagoruyko. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Iclr*, pages 1—-, 2014.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[21] Z. W. TMai Xu, Yun Ren. Learning to predict saliency on face images. In *ICCV*, 2015.

[22] J. W. Y. Y. F. Z. Xiao Liu, Tian Xia and Y. Lin. Fully convolutional attention networks for fine-grained recognition. In *arXiv:1603.06765v4*, 2017.

[23] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[24] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[25] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013. *Computer Vision–ECCV 2014*, 8689:818–833, 2014.

[26] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation Backprop. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9908 LNCS:543–559, 2016.

[27] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

[28] X. W. X. T. Ziwei Liu, Ping Luo. Deep learning face attributes in the wild. In *ICCV*, 2015.