Interpretable Inference Graphs for Face Recognition

Siddhant Garg[*] Goutham Ramakrishnan[*] Varun Thumbe
sidgarg, gouthamr, thumbe @cs.wisc.edu
University of Wisconsin-Madison
[*]Equal contributions by authors

***Abstract*—Convolutional Neural Networks with Adaptive Inference Graphs (ConvNet-AIG) use adaptive network topologies through on/off gating on network layers for individual images to achieve improved computational efficiency and classification accuracy. Face recognition is a more difficult task than object classification due to fine-grained differences in facial features. We evaluate the performance of ConvNet-AIG for the task of Face Recognition on the IMDb-Face dataset. We analyse the interpretability of inference graphs for different facial features and show that facial features like gender, skin color and race can be interpreted corresponding to individual images. We propose a novel loss function to force interpretable inference graphs(ConvNet-IIG) and empirically show an improvement in classification accuracy.**

***Index Terms*—Interpretability, Face Recognition**

I. Introduction

Convolutional neural networks (CNNs) have achieved state-of-the-art performance in several visual tasks, such as object classification and face recognition. Face Recognition is a well studied problem in computer vision with several recent techniques targeting performance improvements ([1], [4], [10], [12], [16], [22]). Recent efforts [18] to provide large scale noise controlled datasets for face recognition have led to the development of the IMDb-Face dataset derived from the MegaFace and MS-Celeb-1M datasets.

Recent progress in computer vision has been dictated by making the convolutional network models deeper [5]. An increased number of training parameters presents two major challenges: increased computational cost and requirement of large amounts of training data. Convolutional Networks with Adaptive Inference Graphs (ConvNet-AIG) [17] mitigate these problems through selectively choosing active layers for individual images using a hard on/off gating on the model layers. Results on image classification datasets like CIFAR-10 [8] and ImageNet [3] show that ConvNet-AIG can improve the performance of vanilla ResNet at the same time reduce computational cost by reducing the number of floating point operations per second (FLOPS).

Model interpretability is another crucial issue for neural networks besides the discrimination power, as discussed in [2]. However, interpretability of CNNs has long been a considerable challenge in computer vision research.

ConvNet-AIG presents empirical evidence that the model automatically learns distinct and interpretable inference graphs for different classes of images. Interpreting inference graphs provides insights such as specific layers being more responsible for different image features, fewer active layers required for classifying 'easy' images, etc. Since every face image has the same physiological features, the task of face recognition is a more fine-grained and difficult task than object-image classification. Facial features like facial hair, ethnicity, skin tone, etc. corresponding to the face images impact the task of classifying face images corresponding to different individuals. A natural question to ask thus is: *Does the improved performance and interpretability of the ConvNet-AIG model, previously well demonstrated for the task of object classification, extend to the task of face recognition?*

The approach of ConvNet-AIG proposes a natural interpretability of the inference graphs stemming from improved classification performance over vanilla ResNet architectures. We reason about the research potential of the reverse of this setting: *Does forcing the learning of interpretable inference graphs based on specific image features lead to an improvement in classification accuracy?* Using interpretable inference graphs for the task of face recognition, the facial features of a test image with no annotations can be estimated.

We use the noise controlled IMDb-Face dataset to evaluate the impact of the ConvNet-AIG framework towards face recognition. We also propose a new loss to force interpretable inference graphs corresponding to facial features and show that this leads to an improvement in the classification accuracy. The major contributions in our paper can be summarized as follows:

- We empirically evaluate the performance of ConvNet-AIGs for the task of Face Recognition on the IMDb-Face dataset
- We provide annotations corresponding to facial features to supplement the information of the IMDb-Face dataset, which can be used by the vision community
- We propose a new loss term (L_{IIG}) to force interpretable inference graphs(ConvNet-IIG) and show improvements in recognition accuracy
- We provide qualitative analysis of the inference graphs by interpreting them vis-à-vis facial features

The rest of the paper is structured as follows: first, we review the related work in the areas of Face Recognition and interpretability. We then succinctly summarize the ConvNet-AIG model and formally define the loss corresponding to interpretable inference graphs (ConvNet-IIG). We go on to

978-1-7281-4187-9/19/$31.0©2019 IEEE

describe the details of the dataset and experimental results and end with future work directions and conclusion.

II. Related work

Seminal works in Face Recognition use deep CNNs trained to optimize embeddings corresponding to faces either by using triplets of roughly aligned matching and non-matching face patches [12] or by applying a piecewise affine transformation using explicit 3D face modeling [16]. [10] proposes an angular softmax loss to angularly discriminative features corresponding to a prior that faces lie on a hypersphere manifold. Recent works use a variety of novel techniques to improve the state of the art like using an additive angular margin loss [4] or using compact fixed-length representations similar to NetVLAD [1] for face recognition [22].

We use residual networks (ResNet) [5] for face recognition by posing it as a classification problem. ResNets combat the vanishing gradient problem by learning residual functions with reference to the layer inputs. Stochastic depth training [7] is used for faster and efficient training of ResNet models by randomly skipping entire layers during training, leading to a shallow back-propagation optimization. ConvNet-AIG [17] models add nuance to the concept of stochastic depth, by incorporating the ability to learn dynamic inference graphs conditional on the input image using gated layers.

ConvNet-AIG [17] models mimic the concept of a hard attention mechanism over the layers of the model similar to related works ([6], [15]) by prioritizing some layers over others. Inference graphs can be analysed similar to visualising attention mechanisms based on importance of layers in classification for different classes.

Recent works focus on interpretability of deep network models for visual tasks. Several works [11], [14], [20] visualize filters in a CNN to explore the pattern hidden inside a neural unit. [21] assign each filter in a high convolutional-layer of a CNN with an object part during the learning process to clarify knowledge representations for CNNs. Related efforts include CAM [23] which leverages global max pooling to visualize dimensions of the representation and Grad-CAM [13] which relaxes constraints on the network with a general framework to visualize any convolution filters. Early methods [9] for interpreting face recognition target to improve the recognition accuracy. [19] propose a spatial activation diversity loss to learn structured face representations which are discriminative and robust to occlusions.

III. Methodology

A. Adaptive inference graphs

ConvNet-AIG [17] use 'gated' layers in a ResNet architecture, allowing execution of inference graphs with dynamic topologies. This reduces the computational cost in terms of number of floating point operations carried out during both training and inference times. The 'gated layer' is defined as:

$$\mathbf{x}_l = \mathbf{x}_{l-1} + z(\mathbf{x}_{l-1}) \cdot F_l(\mathbf{x}_{l-1}) \tag{1}$$

where $\mathbf{x}_l$ and $F_l(.)$ represent the function computed and the output of the l^{th} layer of the network respectively. The 'gate' is represented by the Bernoulli variable z, whose value is conditional upon the output of the previous layer. The 'gate' is 'open' or 'closed' when z assumes values of 1 and 0 respectively, with the 'closed' gates providing the reduction in computation. Note that the analogous expression for a traditional ResNet architecture would be identical to 1 except for the multiplicative Bernoulli variable z.

The modeling of 'gates' as discrete Bernoulli variables results in a non-differentiable model, and makes learning the 'gates' difficult through traditional gradient methods. To get around this, the authors use the Gumbel-Max trick to relax the discrete gate variable into a softmax which allows for the gate parameters to be learned through back-propagation.

The output of each 'gate' is conditional upon the feature maps obtained from the previous layer. The architecture of each gate consists of a global average pooling layer, followed by two fully connected layers. This output (referred to as the relevance score by the authors) is used for greedy Gumbel sampling, which represents a differentiable approximation to the discrete decision of whether to execute the layer or not. Further details can be found in Section 3 of [17].

In theory, each convolutional layer can have its own 'gate'. However, the authors of ConvNet-AIG consider each individual block of the ResNet as the atomic element of the inference graph. This reduces the total number of 'gates' in the deep network, simplifies the architecture and thus improves interpretability. We continue to follow this convention in our implementation and experiments.

ConvNet-AIGs achieve reduced computations by encouraging each layer to 'execute' (i.e. have an 'open' gate) for only a fraction of images in each mini-batch, referred to as the 'target rate'. This is done by introducing an additional term in the training loss:

$$L_{target} = \sum_{l=1}^{N} (\bar{z}_l - t_l)^2 \tag{2}$$

where $\bar{z}_l$ is the fraction of images in the minibatch for which layer l was executed (obtained by simply averaging over the values of the gates for each image) and t_l is the predefined target rate. The target loss is summed mean squared error between the desired and achieved target rates over all layers of the network. The composite training loss is the sum of the classification (cross-entropy) loss and the target loss.

$$L_{train} = L_{CE} + L_{target} \tag{3}$$

B. Interpretable inference graphs

The experimental evaluation of the ConvNet-AIG model on the ImageNet dataset showed that the network learned distinct inference graphs for different classes of images. Thus along with reduced computations, the use of adaptive inference graphs also resulted in improved interpretability and understanding of the model. We take this idea further by proposing ConvNet-IIG (Interpretable inference graphs).

We explicitly encourage interpretability of the layers of the neural network, by incorporating auxiliary information about the dataset through a novel loss term.

In several classification tasks, one may have access to auxiliary information about the data in the form of additional annotations, in addition to the images at training time, but not during inference. For example - for the task of face recognition, one may have information about specific facial features such as eye color and skin tone. Interpretability of the layers of the model in terms of this auxiliary information would be useful and desirable. To achieve this we propose a modified composite training loss, replacing the target loss of the ConvNet-AIG with a novel loss for the ConvNet-IIG, which is defined below.

Let $G = [G_1, \ldots, G_k]$ denote the gates in the model and $C = \{C_1, \ldots, C_p\}$ denote the p possible labels assumed by the annotation feature $\mathbb{F}$ for each image. Choose p disjoint subsets of $G : \{M_{C_1}, \ldots, M_{C_p}\}$, to denote the mask corresponding to the desired active gates for each distinct value C_i that the feature can take (Disjoint subsets are necessary to avoid competing terms in the overall loss). Let $\{X_1, \ldots, X_N\}$ be the set of training images in the mini-batch, and $\{c_1, \ldots, c_N\}$ denote the annotations for each image. Let $\{g_{i1}, \ldots, g_{ik}\}$ denote the binary values representing the executed gates for input image X_i. Then the loss term is defined as:

$$L_{IIG}(\mathbb{F}) = \sum_{i=1}^{N} \sum_{j=1}^{k} (\mathbb{1}_{\{G_j \in M_{c_i}\}} - g_{ij})^2 \tag{4}$$

where $\mathbb{1}_{\{G_j \in M_{c_i}\}}$ is 1 if $G_j \in M_{c_i}$ and 0 otherwise.

During inference, we can use ConvNet-IIG to estimate facial features for test images with no annotations. Intuitively, we map a test image X to a label C_p for a particular facial feature $\mathbb{F}$ if the active layers in ConvNet-IIG for X are the same as those for that label C_p. Mathematically, let $\{g_1, \ldots, g_k\}$ denote the binary values representing the executed gates for a test image X through a trained ConvNet-IIG model. The label of the annotation feature $\mathbb{F}$ for X can be estimated as:

$$\mathbb{F}(X) = C_p \text{ where } p = \underset{i}{\operatorname{argmin}} \sum_{j=1}^{k} (\mathbb{1}_{\{G_j \in M_{c_i}\}} - g_j)^2 \tag{5}$$

IV. Evaluation

We investigate the performance and interpretability of the ConvNet-AIG and ConvNet-IIG models for the task of face recognition. First, we describe the dataset used for the evaluation and the methodology used for obtaining additional annotations. Then, we describe qualitative and quantitative results from experiments.

A. Dataset

The IMDb-Face dataset [18] is a new large-scale noise-controlled dataset for Face Recognition made by cleaned subsets of popular face databases.

- The complete dataset has around 59k different celebrity face identities, but is unbalanced since not all different face identities have the same number of images. In order to enable the qualitative study of different network layers learning different facial features, we heuristically choose to prune the dataset to only the faces with at least 100 images in order to balance the training data.
- We restrict ourselves to the task of face recognition, and therefore use the bounding box information to crop out the images of the faces. We scale all images to a fixed size of $[128 \times 128]$.
- Some facial images have highly distorted image aspect ratios (especially in cases where the image corresponds to the side poses of the face). To avoid the skewing of images due to rescaling, we heuristically prune the dataset by retaining only the images whose original aspect ratio is in the range $[0.75, 1.25]$.

TABLE I
Additional annotations of facial features

Facial Feature	Values
Gender	Male, Female
Skin Color	Black, Brown, White
Eye Color	Black, Brown, Blue
Facial Hair	None, Beard, Moustache, Both Beard and Moustache
Race	African Descent, Asian(Chinese), Asian(Indian), Middle-Eastern, Latino, Caucasian, Other
Age	less than 35, 35-45, 45-55, greater than 55

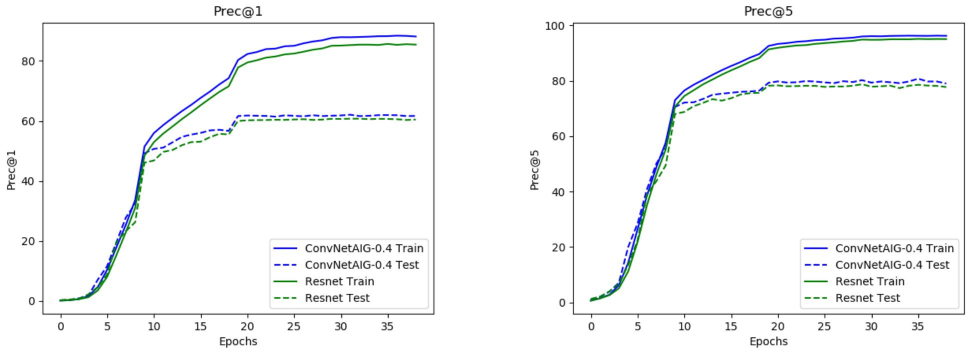
There are no predefined train/test data splits available for IMDb-Face, thus we construct our dataset along the specifications detailed above. Our final dataset consists of 100 face images for each of the 750 celebrities which are further randomly split into 80 train and 20 test images per celebrity.

Additional Annotations In [17], ConvNet-AIG learns distinct inference graphs for different classes of the ImageNet dataset. However, the variance between the different images classes of IMDb-Face is comparatively smaller, as the images are of human faces. We examine if ConvNet-AIG can learn distinct inference graphs for the subtle facial features such as gender, ethnicity and eye color. The proposed ConvNet-IIG model based on the L_{IIG} loss term also uses the facial features as the auxiliary information for forcing interpretability.

We use web-scraping and manual annotations to compile the auxiliary information for 6 facial features (gender, skin color, eye color, ethnicity/race, facial hair and age) for each celebrity face in our dataset. Here, we chose these specific features since they lend themselves to easy interpretation by humans. The details of different labels for each facial feature are shown in Table I. For manual annotations of features such

TABLE II
Quantitative Results

Model	Top-1 Test Accuracy	GFlops
ResNet-101	60.69	2.54
ConvNet-AIG [t=0.4]	62.07	1.27
ConvNet-AIG [t=0.5]	60.85	1.75
ConvNet-AIG [t=0.7]	61.15	2.06
ConvNet-IIG (Gender)	62.41	1.73
ConvNet-IIG (Gender + Age)	62.29	1.18



(a) Top-1 accuracy

(b) Top-5 accuracy

Fig. 1. The train and test curves for Top-1 and Top-5 accuracy of the ConvNet-AIG model. Best viewed in color.

as skin color, we used a majority voting scheme between labels from different annotators [1]. An unknown label is used for each feature in the case where its not applicable/relevant(for example, age for deceased celebrities), and these were labels were not considered for the qualitative analysis.

B. Experimental Details

We use a ResNet-101 architecture for our experiments, with the final average-pool layer modified appropriately to account for our input image size. The model has 33 BottleNeck blocks, where each block contains 3 convolutional and 3 batch-normalization layers. Previous experiments on the ConvNet-AIG model use a target rate $t_l = 1.0$ for the first 7 and last 3 blocks, arguing that low-level and high-level feature computations must be common for all images. We adopt the same experimental setup, having only the middle 23 Bottleneck blocks with 'gates' (over which L_{target} and L_{IIG} are computed). All models were trained with a batch-size of 150, on a 12GB NVIDIA Tesla K80 GPU.

C. Quantitative Results

The quantitative experimental results comparing native ResNet-101, ConvNet-AIG and ConvNet-IIG models are summarized in Table II. We present the best test accuracy for each model architecture over multiple runs. The ConvNet-AIG model was trained with three different target rates (0.4, 0.5 and 0.7 respectively for the middle 23 blocks, with a 1.0 target rate for the first 7 and last 3 blocks).

We train two ConvNet-IIG models by incorporating the new loss term L_{IIG} described in Section III. Similar to experiments with ConvNet-AIG, we set the target rate t_l to 1 for the first 7 and last 3 blocks. We use a simple heuristic scheme for choosing the masks $\{M_{C_1}, \ldots, M_{C_p}\}$ corresponding to the facial feature labels. The first model is trained using gender as the auxiliary information feature with alternating blocks (from the 23 blocks) masking the male/female labels. The second model uses gender and age as the auxiliary information features. Of the 23 middle blocks, we use 3 blocks to mask the male/female gender labels each and 4 blocks to mask each of the 4 brackets corresponding to age.

All three ConvNet-AIG models achieve a better classification accuracy than the native ResNet model, corroborating the results obtained in [17]. Note that the ConvNet-AIG models have more trainable parameters (due to the parameters in the 'gates'). However, the advantage of the ConvNet-AIG model is seen in terms of the fewer GFlops (Giga-Floating Point Operations Per Second) during model training. The model with the lowest target rate(0.4) achieves the best classification accuracy among the three. We empirically find no visible correlation between target rate t_l and the model test accuracy. This may be attributed to the fact that the feature representations learnt by a given layer of the model will depend on the corresponding gate function, which in-turn depends on the target rate. For lower target rates, a smaller number of active layers necessitates each layer to learn more discriminative features. However, we also observe that the models converge to roughly the same Top-1 accuracy for different target rates. We present graphs to show how the Top-1 and Top-5 accuracy changes over training the ConvNet-AIG [t=0.4] on the train and test data in Figure 1.

Both the ConvNet-IIG models marginally outperform the native ResNet and ConvNet-AIG models in Top-1 classification accuracies. This suggests that incorporation of the auxiliary information through the new loss term L_{IIG} is

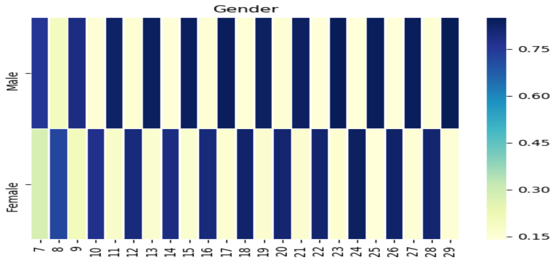


Fig. 2. The ConvNet-IIG (Gender) heatmap for Gender

[1] We acknowledge the subjectivity in the annotated features including the choice of discrete labels for each feature. Our annotations are based upon a customary notions of human facial features.

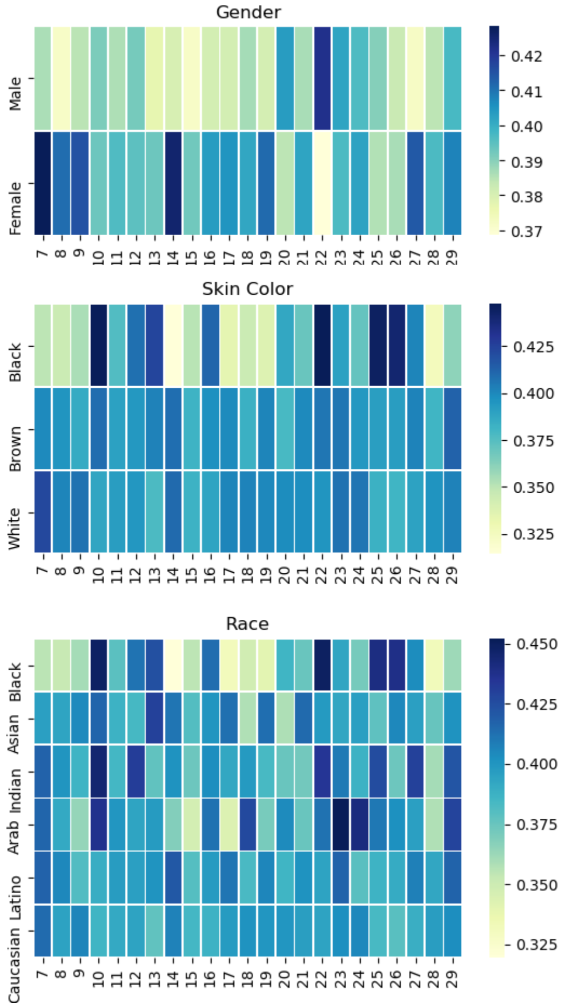


Fig. 3. The ConvNet-AIG[t=0.4] heatmaps for Gender, Skin Color and Race

beneficial. However, the additional utility of the ConvNet-IIG model stems from the fact that it can be used to annotate feature information for images at inference time as described in Section III. An interesting observation is the marginal drop in test accuracy when adding age as the auxiliary feature to gender. This may be due to one of two reasons: age may not be a good feature for face recognition, or our heuristic for masking gates when computing L_{IIG} is not well crafted. The GFLOPS for the ConvNet-IIG (Gender + Age) is smaller than for the ConvNet-IIG (Gender) model. This is a direct consequence of our choices of $\{M_{C_1}, \ldots, M_{C_p}\}$ for the two models, which always activates 7 gates out of the middle 23 in the former compared to at least 11 in the latter.

D. Qualitative Results

For each image in the test set, we plot inference graphs corresponding to which gates of the model were active. The results are summarized below.

1) Heatmaps: We create heatmaps using additional annotation features on the dataset, to examine whether the model is able to learn and distinguish common facial features. The average fraction of images which activate the 'gate' of a given block for a particular label of the facial feature are quantified using different colors.

For the ConvNet-IIG model, the heatmaps correspond to the masking heuristic used for the L_{IIG} loss. The heatmap for 'Gender' obtained from the ConvNet-IIG (Gender) model shown in Figure 2 displays the alternating scheme of activated gates used in the design of the loss function. This demonstrates the utility of the model at inference time, with the gate activations providing insights into interpretable facial features along with the classification of a test image.

We now provide qualitative analysis of heatmaps obtained from the ConvNet-AIG [t=0.4] model. We choose this specific target rate since (a) it achieves the best classification accuracy among the three ConvNet-AIG models and (b) we observe that heatmaps from models with smaller target rates show greater variance, and thus greater scope for interpretability in the resulting inference graphs.

In Figure 3, we plot three heatmaps which show significant inter-class variance. From the gender heatmap, one can see that some layers are activated more for one gender over the other (like layer 14 for females and layer 22 for males) showing that specific layers learn the feature corresponding to different genders. Similar variance in gate activations can be observed from the heatmaps for Skin Color and Race. These observations suggest that these facial features are important

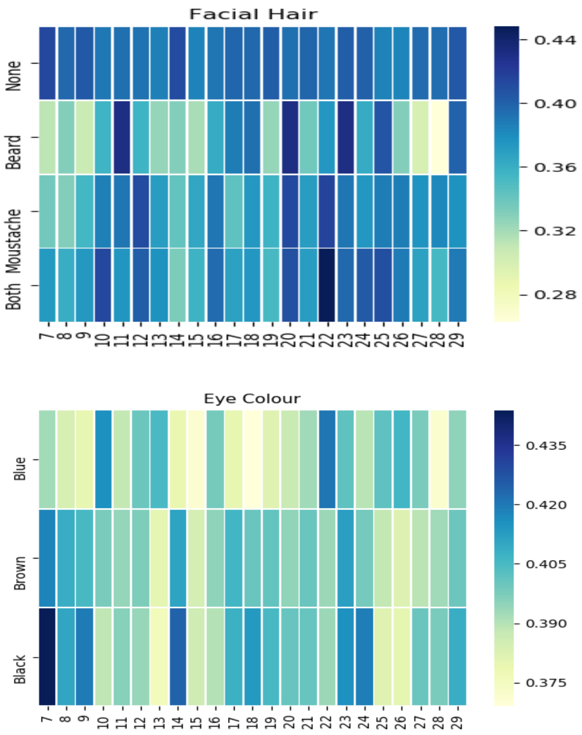


Fig. 4. The ConvNet-AIG[t=0.4] heatmaps for Facial Hair and Eye Color

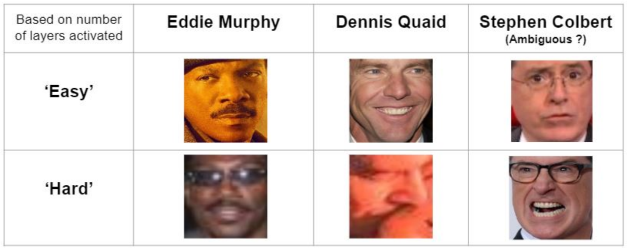


Fig. 5. Analyzing ease of classification of images of the same class based on number of active layers used

for the task of face recognition.

However, not all heatmaps obtained were as informative. In Figure 4, we show the heatmaps corresponding to the Facial Hair and Eye Color. These are more difficult to interpret and suggest that these features are less important for the task of face recognition. This is also intuitive, as facial hair and eye color are more subtle features to detect than skin color or gender and thus may be harder for the model to detect.

2) Easy/Hard Images: An analysis of the 'difficulty' of classification of different images of the same class was presented in [17], on the basis of the number of layers which were active in the model (with fewer layers corresponding to 'easier' images and vice-versa). We perform an analogous analysis for the task of face recognition, and observe intra-class variations in the number of active gates for different images. From Figure 5, it can be seen that the model found it 'easier' to classify some images than others (all images shown were correctly classified). For celebrities Eddie Murphy and Dennis Quaid, the easy and hard images are intuitive, i.e. the image the model found to be hard is also hard for a human to classify. However, the example images of Stephen Colbert show that this might not always be the case; the relative difficulty of classifying the two images is ambiguous. Another interesting observation we make in this analysis: very few layers of the model are active when the model incorrectly classifies an image, often skipping all the layers which it can.

V. Future Work

Future work for using ConvNet-AIG for Face Recognition is to investigate the interpretability of the inference graphs corresponding to facial features for face recognition datasets from other domains. In this paper we only study a simple strategy to force interpretable inference graphs using alternate layers. A future direction of work can be to learn strategies jointly along with the model weights using a probability distribution over the layers of the model. Studies suggest the improvements in classification accuracy from transfer learning pre-trained weights for other tasks. A natural question thus, is to study if the interpretability of inference graphs holds for the case of using a pre-trained model.

VI. Conclusion

In this paper we empirically evaluate the performance gains of using ConvNet-AIG for the task of Face Recognition. We show that inference graphs can be interpreted easily for facial features like gender and skin color but are more difficult to do so for features such as facial hair and eye color. We develop a technique for forcing interpretable inference graphs and show that it leads to an improvement in classification accuracy. Using this technique, we provide a means for estimating feature annotations for test images. We also release annotations of facial features corresponding to the IMDb-Face dataset for advances in facial recognition by the computer vision community.

References

[1] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *CoRR*, abs/1511.07247, 2015.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *CoRR*, abs/1704.05796, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009.

[4] Jiankang Deng, Jia Guo, and Stefanos P. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CoRR*, abs/1801.07698, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[7] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. *CoRR*, abs/1603.09382, 2016.

[8] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[9] Erik Learned-miller, Gary Huang, Aruni Roychowdhury, Haoxiang Li, Gang Hua, Erik Learned-miller, Gary B. Huang, Aruni Roychowdhury, and Haoxiang Li. Labeled faces in the wild: A survey.

[10] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *CoRR*, abs/1704.08063, 2017.

[11] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *CoRR*, abs/1412.0035, 2014.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[13] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[14] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[15] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.

[16] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] Andreas Veit and Serge J. Belongie. Convolutional networks with adaptive computation graphs. *CoRR*, abs/1711.11503, 2017.

[18] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. *CoRR*, abs/1807.11649, 2018.

[19] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. *CoRR*, abs/1805.00611, 2018.

[20] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[21] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. *CoRR*, abs/1710.00935, 2017.

[22] Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. Ghostvlad for set-based face recognition. *CoRR*, abs/1810.09951, 2018.

[23] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.