Visualizing Bias in Activations of Deep Neural Networks as Topographic Maps

Valerie Krug[1,*], Christopher Olson[1] and Sebastian Stober[1]

[1]*Artificial Intelligence Lab, Otto-von-Guericke-University Magdeburg, Germany*

Abstract

Deep Neural Networks (DNNs) are successful but work as black-boxes. Elucidating their inner workings is crucial as DNNs are prone to reproducing data biases and potentially harm underrepresented or historically discriminated demographic groups. In this work, we demonstrate an approach for visualizing DNN activations that facilitates to visually detect biases in learned representations. This approach displays activations as topographic maps, similar to common visualization of brain activity. In addition to visual inspection of activations, we evaluate different measures to quantify the quality of the topographic maps. With visualization and measurement of quality, we provide qualitative and quantitative means for investigating bias in representations and demonstrate this for activations of a pre-trained image recognition model when processing images of peoples' faces. We find biases for different sensitive variables, particularly in deeper layers of the investigated DNN, and support the subjective evaluation with a quantitative measure of visual quality.

Keywords

explainable AI, deep neural networks, topographic activation maps, representation analysis

1. Introduction

Deep Neural Networks (DNNs) are highly successful but it is difficult to interpret how they perform their learned task [1]. This is particularly dangerous in critical applications where biases in the decision making can negatively affect certain groups of people, often those who are underrepresented and discriminated against already. To detect undesired behavior of DNNs, model introspection aims to better understand their inner processes. In this work, we investigate biases in representations of DNNs. In particular, we visualize activations as topographic maps, similar to how brain activity is commonly presented [2]. We explain the visualization approach and evaluate different measures of visual quality of the topographic maps. Then, we use our technique to investigate representational bias in a pre-trained image recognition model.

Introspection Feature visualization explains learned patterns by creating inputs that maximally activate particular filters [1, 3, 4]. Attribution techniques explain the output of a DNN for input examples by quantifying the relevance of each input value for the output [5, 6, 7, 8]. Data representation analysis investigates activations of a large amount of data [9, 10, 11, 12, 13],

Aequitas 2023: Workshop on Fairness and Bias in AI | co-located with ECAI 2023, Kraków, Poland

*Corresponding author.

valerie.krug@ovgu.de (V. Krug); christopher.olson@ovgu.de (C. Olson); stober@ovgu.de (S. Stober)

0000-0002-4729-1840 (V. Krug); 0000-0002-1717-4133 (S. Stober)

some provide a graphical user interface [14, 15, 16, 17] and some investigate training processes [18]. Our approach also analyzes and visualizes representations but focuses more on the ease of visual inspection and less on highly detailed information about activation similarity of neurons.

Bias Detection and Mitigation DNNs are prone to reproducing or emphasizing biases of data they are trained on. Different approaches to detect and mitigate bias have been introduced. For example, researchers showed racial discrimination in the online ad delivery by Google [19], debiased word embeddings [20] or evaluated discrimination in facial recognition [21]. More recently, researchers investigated biases in transformer-based models [22, 23, 24]. Balanced evaluation data sets like Gender Shades [21] or Fair Face [25] facilitate bias analyses. In this work, we visualize bias in representations of a DNN, different to investigating its output [21, 25].

2. Method

In this section, we introduce our approach of visualizing DNN activity as topographic maps. An implementation is available at https://github.com/valeriekrug/ANN-topomaps.

Group-Specific Activations We use an averaging approach to characterize DNN activity for groups of examples [26]. For each group, we average the activations in the layer of interest and subtract the average over all groups. We obtain positive and negative values that represent higher and lower activity in comparison to the other groups. Finally, we stack them for all groups as a $G \times N$ matrix, where G and N denote the number of groups and neurons, which we refer to as the Neuron Activation Profile (NAP). Notably, any grouping can be used, independent of the predicted model classes.

Topographic Activation Map Inspired by how brain activity is displayed as topographic maps, we map DNN neurons to allow for a similar activation visualization [27]. First, we distribute neurons in a 2D space such that neurons of similar activity are close to each other with a UMAP projection. Then, we evenly distribute the neurons in the 2D space by treating them as particles that attract each other to close gaps and repel others to avoid two particles at the same position. For the set of particles P, we compute a force for each particle $i \in P$

$$f(i) = \frac{\sum_{j \in P \setminus i} (attr(i,j) - rep(i,j))}{|P \setminus i|}$$
$$attr(i,j) = 1.5 \cdot (dist(i,j) + 1)^{-3} \qquad rep(i,j) = 15 \cdot e^{-(dist(i,j)/2)} \tag{1}$$

where $dist$ = Euclidean distance of particle coordinates and apply it for 1000 iterations.

Finally, we visualize the NAP in the computed layout, mapping the values to a 0-symmetric continuous color scale from blue over white to red. This way, equal colors represent the same value in each group. Then, we linearly interpolate the colors. In Convolutional Neural Networks (CNNs), we compute the layout such that each feature map is assigned a position in a 2D space and that similar feature maps are close to each other. To assign each position a color for a group, we use the respective feature map's mean NAP value.

Quality Measures We investigate approaches to measure the visual quality of topographic activation maps in terms of ease of visual interpretability. We consider the quality as high if there are few distinguishable regions which jointly cover a large area. This does not imply quality of the activations themselves but only the visual quality. Figure 1 shows relevant steps of computing the different measures.

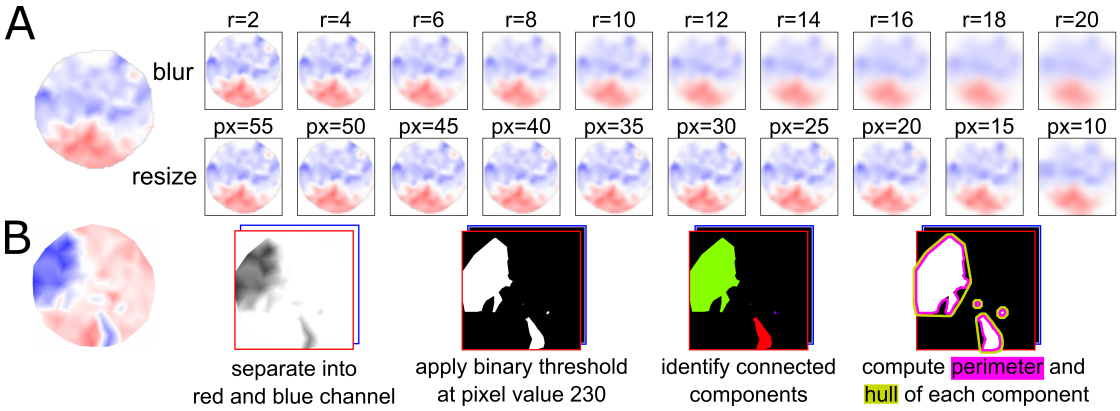

Figure 1: Pre-processing of topographic activation map images for quality measure computation.

To test whether each position in a topographic map is similar to its neighborhood, we measure robustness against image perturbations. We perturb either by a Gaussian blur or by downscaling and then upscaling to the original size (with bicubic interpolation). Then, we compute the Mean Squared Error (MSE) between the perturbed and the original image. We use Gaussian blur with radii 2 px to 20 px in steps of 2 px and investigate downscaling sizes to 55×55 px to 10×10 px in steps of 5 px (see Figure 1A). Finally, we aggregate the MSEs for the different parameters with an estimated area under the curve (AUC) value (trapezoidal rule). We call the measures "blur MSE AUC" and "resize MSE AUC".

We further quantify topographic map quality based on connected components (compare Figure 1B). First, we separate the image into the red and blue channel. For both channels, we apply a binary threshold at pixel value of 240 to separate regions from the background. Note that small values in the red channel indicate a blue region and vice versa. In the binarized images, we detect connected components using OpenCV[1]. We compute the number of components larger than 10 px area ("count") and the average component size relative to the circle area.

Generally, few large components are considered as high quality. However, this does not account for whether the components are large but interwoven with others. Therefore, we further compute the convexity of each connected component as h/p, where h is the length of the convex hull and p the perimeter of the connected component (Figure 1B on the right depicts perimeter and hull). Finally, we aggregate convexity values of the components and reward larger components. To this end, we compute the fraction of total circle area occupied by each component and use these as weights for a weighted sum of convexity values. This results in a value in the range of $[0, 1]$. We will refer to this quality measure as "size-weighted convexity".

[1]https://github.com/opencv/opencv-python

3. Evaluating Quality Measures

3.1. Experimental Design

For evaluation of the measures, we use a simple model and data set. As data, we use MNIST [28]. MNIST contains grayscale images of handwritten digits from 0 to 9, which are of size 28×28 px, centered and normalized in scale. There are $60,000$ training and $10,000$ test data examples. We train a Multi-Layer Perceptron (MLP) with one fully-connected hidden layer of 128 neurons and ReLU [29] activation. During training, we use dropout with a dropout rate of 0.5. We use TensorFlow [30], with batch size of 32 for 1 epoch, Adam optimizer [31] with default parameters and categorical cross-entropy. Our evaluation uses the MNIST test data set and activations from the hidden layer. Based on a set of manually created topographic activation maps, we choose a quality measure that best describes the visual quality under different conditions. We then use this measure for the representation bias experiment in Section 4.

3.2. Results

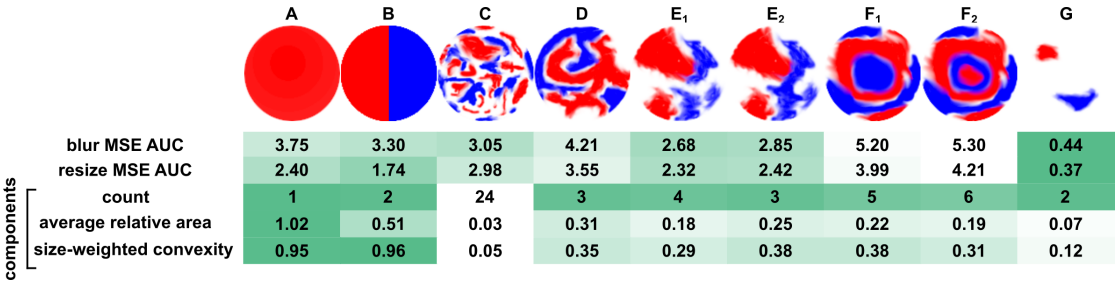

Figure 2: Comparison of different quality measurements on manually created topographic activation maps. Better scores for each quality measure are indicated by darker shades of green.

We manually created the topographic activation maps shown in Figure 2 as representative cases to see which quality measure aligns best with our expectations. Maps A and B represent ideal topographic activation maps for which we expect the highest quality. C represents a poor activation map that is difficult to visually interpret, so its quality should be low. D has few large components, however, they are interwoven with each other and should have lower visual quality. E_1 and E_2 are more realistic and visually qualitative examples which differ in that the blue region is split by a small gap. Both should obtain good visual quality and the gap should not affect the value too strongly. F_1 and F_2 are examples to test whether nested components are separated correctly. For our last example G, which represents sparse regions, we expect a low quality that should be higher than that of C as it contains more information.

Comparing the measures, size-weighted convexity is most consistent with our expectations as it gives the highest quality to A and B and the lowest to C and G. Only the low convexity of D in comparison to $E_{1,2}$ or $F_{1,2}$ does not reflect well. Therefore, we will continue using the size-weighted convexity quality measure in the following. Note that average relative component size and size-weighted convexity are not 1 or 0.5 for the ideal examples A and B due to the pixel grid and approximations when obtaining connected components.

4. Bias Analysis

4.1. Experimental Design

We perform the bias experiments for representations of VGG16 [32], a pre-trained CNN model that can be used as feature extractor for downstream applications like image recognition DNNs. As test data, we use FairFace [25], a balanced data set of images of people from different age groups, races and binary genders. Moreover, to investigate the significance, we compare to groups of randomly drawn examples. We obtain VGG16 from TensorFlow Keras applications[2] module and use the second and fifth maxpooling layer (layers 6 and 18) as an example.

4.2. Results

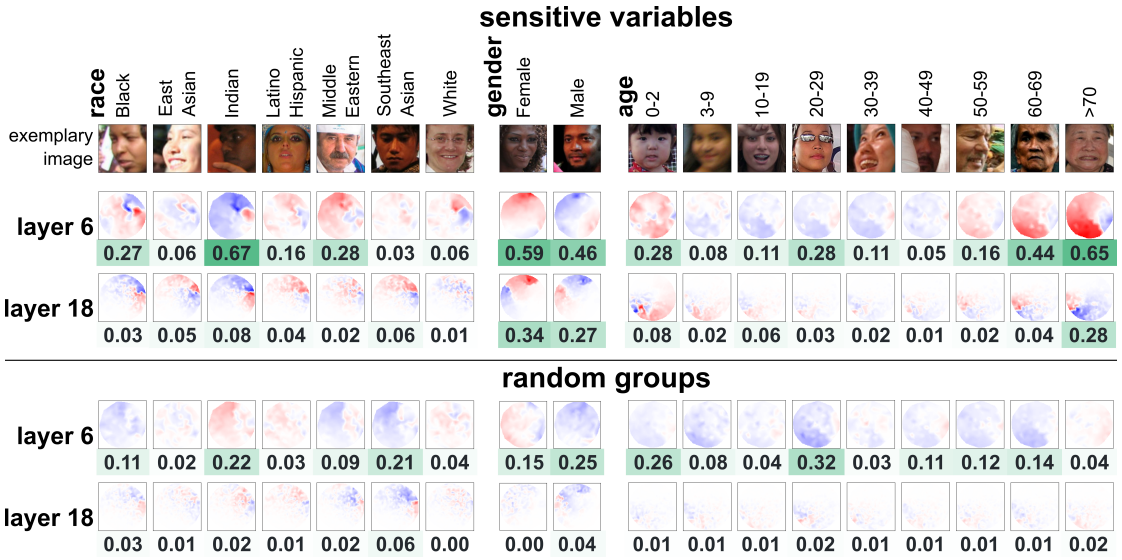

Figure 3: Topographic activation maps when grouping by different sensitive variables in layer 6 and 18 of VGG16 and when grouping randomly. Topographic maps that belong to the same sensitive variable and to the corresponding random groups use a common color map in the same layer. Convexity quality is shown below each topographic map.

Topographic maps for sensitive variables in layers 6 and 18 of VGG16 are shown in Figure 3. Appendix Figure 4 shows results for several more layers of in VGG16.

"race": In layer 6, each category has a specific activation pattern. However, in layer 18, class activations become more similar between particular groups. Specifically, Black and Indian categories are highly similar, as well as East Asian, Southeast Asian and Latino Hispanic. The observations indicate that there is a racial bias in deeper layers. Surprisingly, the Middle Eastern and White categories do not show clear activation patterns, potentially because the model learns more individual representations for these categories.

[2]https://github.com/keras-team/keras

"gender": Topographic activation maps of Female and Male category almost are the inverse of each other, which is expected for a binary grouping. Clearly, in layer 18, the Female and Male groups show stronger over-/under-activation than the random groups, indicating that the representation in deeper layers distinguishes between these categories.

"age": Groups of similar age are similarly activated, which is reasonable considering the categories' fuzzy boundaries. In layer 6, we observe clusters of high similarity: age groups 0-2, 10-49 and >50. Layer 18 shows more continuous changes and strongest activation deviation from the mean in the lowest and highest age groups. There seems to be no systematic disadvantage for any individual group but the groups would be distinguishable in a downstream application.

Significance To evaluate the significance of the results, we contrast sensitive variables and random groups regarding color intensity of topographic maps and visual difference of groups.

We observe stronger color intensity for sensitive variables than random groups. Further, activation differences are more pronounced between the sensitive variables than between the random groups. Both indicate that the observed similarities and patters are not only a random effect but really related to the sensitive variables.

We notice that visual quality is generally lower in layer 18. This might be related to the higher number of feature maps in layer 18 (512) compared to layer 6 (128). Visualizing more feature maps means that each influences a smaller part of the map. Therefore, the regions are likely to become less convex. We further observe that there is a large white region common to all groups, which decreases the visual quality. It represents feature maps that are inactive or unspecific to the groups but potentially sensitive to features that are not present in face images.

In general, the quality of visualizations for sensitive variables is higher than for the corresponding random groups which supports the significance of the results.

Diversity of Groups We appreciate the efforts of the FairFace data set to provide balanced evaluation data. However, the sensitive variables still have potential to be further diversified. For example, "gender" is only considered as a binary or age group ">70" includes a larger range of ages than other groups. We still consider the data suitable to demonstrate our technique but encourage the community to conduct studies with more diverse data sets upon availability.

5. Conclusion

Topographic activation maps are a promising tool to visually inspect bias in representation of DNNs and our visual quality measure supports the otherwise only subjective evaluation.

Our approach does not provide an explanation of the patterns responsible for the bias, neither do we mitigate biases. We consider our method to be a visual overview to spot likely biases to look for. Moreover, we expect that topographic activations maps are useful for people without expert knowledge in Machine Learning to get a simplified insight into DNN internals.

In this work, we only considered the sensitive variables independently. As this does not consider intersectionality, we will investigate combinations of sensitive variables in future work.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany (BMBF) as part of the project "CogXAI – Cognitive neuroscience inspired techniques for eXplainable AI". We also thank Raihan Kabir Ratul for implementing FairFace dataset pre-processing.

References

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

[2] K. Maurer, T. Dierks, Atlas of Brain Mapping: Topographic Mapping of EEG and Evoked Potentials, Springer Science & Business Media, 2012.

[3] D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing higher-layer features of a deep network, University of Montreal 1341 (2009) 1.

[4] A. Mordvintsev, C. Olah, M. Tyka, Inceptionism: Going deeper into neural networks, Google Research Blog. Retrieved June 20 (2015) 5.

[5] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision (ECCV), Springer, 2014, pp. 818–833.

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

[7] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution, International Conference on Learning Representations (ICLR) (2018).

[8] K. Schulz, L. Sixt, F. Tombari, T. Landgraf, Restricting the flow: Information bottlenecks for attribution, International Conference on Learning Representations (ICLR) (2019).

[9] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, International Conference on Learning Representations (ICLR), Workshop Track Proceedings (2017).

[10] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), in: International Conference on Machine Learning (ICML), 2018, pp. 2668–2677.

[11] J. Fiacco, S. Choudhary, C. Rose, Deep neural model inspection and comparison via functional neuron pathways, in: Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 5754–5764.

[12] A. S. Morcos, M. Raghu, S. Bengio, Insights on representational similarity in neural networks with canonical correlation, arXiv preprint arXiv:1806.05759 (2018).

[13] T. Nagamine, M. L. Seltzer, N. Mesgarani, Exploring how deep neural networks form phonemic categories, Annual Conference of the International Speech Communication Association (Interspeech) (2015).

[14] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, C. Olah, Activation atlas, Distill (2019).

[15] F. Hohman, H. Park, C. Robinson, D. H. P. Chau, Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations, IEEE transactions on visualization and computer graphics 26 (2019) 1096–1106.

[16] H. Park, N. Das, R. Duggal, A. P. Wright, O. Shaikh, F. Hohman, D. H. P. Chau, Neurocartography: Scalable automatic visual summarization of concepts in deep neural networks, IEEE Transactions on Visualization and Computer Graphics 28 (2021) 813–823.
[17] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (2019) 56–65.
[18] M. Li, Z. Zhao, C. Scheidegger, Visualizing neural networks with the grand tour, Distill 5 (2020) e25.
[19] L. Sweeney, Discrimination in online ad delivery, arXiv preprint arXiv:1301.6822 (2013).
[20] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, volume 29, 2016, pp. 4349–4357.
[21] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81, PMLR, 2018, pp. 77–91.
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
[23] B. Li, H. Peng, R. Sainju, J. Yang, L. Yang, Y. Liang, W. Jiang, B. Wang, H. Liu, C. Ding, Detecting gender bias in transformer-based models: A case study on BERT, arXiv preprint arXiv:2110.15733 (2021).
[24] J. Ahn, A. Oh, Mitigating language-dependent ethnic bias in BERT, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 533–549.
[25] K. Karkkainen, J. Joo, Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1548–1558.
[26] A. Krug, M. Ebrahimzadeh, J. Alemann, J. Johannsmeier, S. Stober, Analyzing and visualizing deep neural networks for speech recognition with saliency-adjusted neuron activation profiles, MDPI Electronics 10 (2021) 1350.
[27] V. Krug, R. K. Ratul, C. Olson, S. Stober, Visualizing deep neural networks with topographic activation maps, in: HHAI 2023: Augmenting Human Intellect, IOS Press, 2023, pp. 138–152.
[28] Y. LeCun, C. Cortes, MNIST handwritten digit database, http://yann.lecun.com/exdb/mnist/, 2010.
[29] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines (2010) 807–814.
[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL: https://www.tensorflow.org/, software available from tensorflow.org.
[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint

arXiv:1412.6980 (2014).
[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

A. Extended Bias Analysis Results

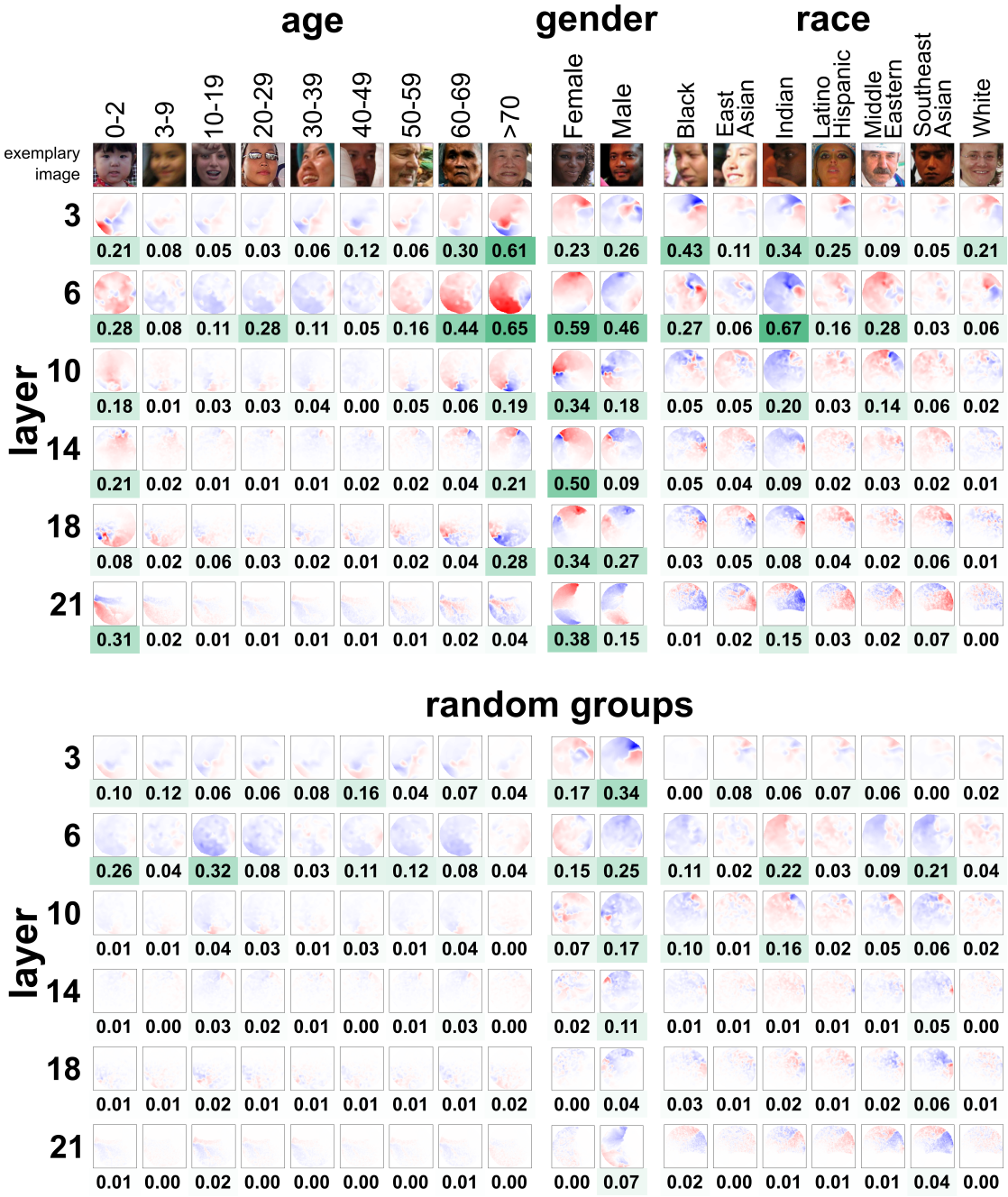

Figure 4: Topographic activation maps when grouping by different sensitive variables in different layers of VGG16 and when grouping randomly. Topographic maps that belong to the same sensitive variable and to the corresponding random groups use a common color map in the same layer. Convexity quality is shown below each topographic map.