Investigating the Interpretability of Biometric Face Templates Using Gated Sparse Autoencoders and Differentiable Image Parametrizations

Peter Rot [1 2] **Klemen Grm** [2]

Abstract

State-of-the-art face recognition models rely on deep, complex neural net architectures that produce relatively compact template vectors, making their mechanisms of operation difficult to interpret and understand. Recently, mechanistic interpretability has emerged as a promising approach to explain large language models. In this paper, we aim to apply such approaches to explain face recognition models. Our method involves transforming face image templates into sparse representations and analyzing their components by identifying images that maximize activation. Our results demonstrate that existing mechanistic interpretability techniques generalize well to previously unconsidered tasks and architectures, and that differentiable image parametrizations can serve as a useful additional means of confirming the interpretation of sparse representations.

1. Introduction

Face recognition models are becoming increasingly vital for high-stakes applications like personal identification and law enforcement. However, as face recognition models become more complex to improve accuracy (Rajpal et al., 2023), their decision-making processes become less understandable. They often operate as black-box models, raising concerns about transparency, especially where precise explanations are necessary. Face recognition templates encode information in a highly entangled manner, where individual features often encapsulate multiple factors of variation, rendering them non-understandable to humans. To better understand and interpret face recognition models and their decisions, the field of eXplainable Face Recognition (XFR) aims to clarify their decision-making processes (Williford et al., 2020).

[1]Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, Slovenia [2]Faculty of Electrical Engineering, University of Ljubljana, Tržaška c. 25, Slovenia. Correspondence to: Peter Rot <peter.rot@fri.uni-lj.si>.

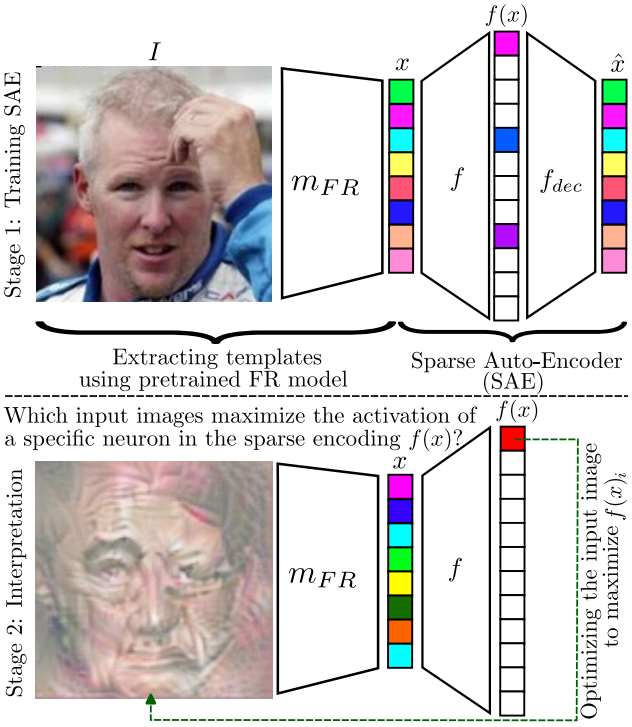

Figure 1. Illustration of our methodology. Given a face recognition model m_{FR} and face images I, we first extract templates $x = m_{FR}(I)$. Templates are then used to train a sparse auto-encoder with an encoder f and decoder f_{dec} to produce sparse representations $f(x)$. In the second stage, we investigate the interpretability of components $f(x)_i$ by optimizing input images to maximally activate individual components of the sparse encoding.

While the interpretability of face recognition models can be approached from different angles, in this work, we draw inspiration from the recent success of *mechanistic interpretability* in explaining large language models (Bricken et al., 2023; Bereska & Gavves, 2024). The high-level overview of our approach is illustrated in figure 1. The idea is to visually inspect which input images maximize the activation of a certain sparse component obtained from a template x. Instead of considering a feature in the original template space x, known to encode information in a highly entangled, compressed manner, we first expand the templates into a sparse representation $f(x)$, obtaining a more decomposed representation. In the first stage, we train a sparse auto-encoder (SAE) on the original templates x to

produce sparse representations $f(x)$. In the second stage, we visualize which input images maximize the activation of specific neurons in these sparse representations. By comparing the optimized visualizations with dataset images that maximally activate a given component of the sparse visualization, we are able to interpret some components of the highly polysemantic face templates, and show that interpretability methods developed for large language models generalize well to different modalities and domains.

2. Related Work

2.1. Mechanistic interpretability

Techniques for *mechanistic interpretability* are an emerging subgroup of explainable artificial intelligence (XAI) techniques that aim to explain the decision making of models by analyzing their underlying algorithms (Bereska & Gavves, 2024; Conmy et al., 2023). While the majority of the literature applies mechanistic interpretability to language models (Bricken et al., 2023; Rajamanoharan et al., 2024), vision models have been proportionally less studied using such methods (Palit et al., 2023).

2.2. Sparse auto-encoders for interpretability

Sparse auto-encoders turned out to be useful for obtaining interpretable features directions in large language models (Cunningham et al., 2023). In (Bricken et al., 2023), the authors decompose features obtained by a language model by expanding them into a sparse representation of a higher dimension using a sparse auto-encoder. They analyze which passages from a large text corpus activate individual components in the sparse representation. In this work, we adapt this approach for a computer vision task, specifically face recognition.

3. Methodology

High-level overview. The high-level overview of our interpretation methodology is presented in Figure 1. Using the state-of-the-art SwinFace (Qin et al., 2023) pretrained face recognition model m_{FR} and the VGGFace2 (Cao et al., 2018) face image dataset, we extract templates $x = m_{FR}(I)$, where I is the input face image. We then use a sparse autoencoder to transform the image templates into sparse encodings $f(x)$, which encode the same information in a more decomposed manner. We look for images that maximally activate a sparse encoding component $f(x)_i = f(m_{FR}(I))_i$ by solving the constrained optimization problem

$$I_{opt} = \arg\max_I f\left(m_{FR}\left(I\right)\right)_i, \quad s.t.\, I = DIP(z) \tag{1}$$

using gradient ascent over I while the parameters of the sparse encoder f and the face recognition model m_{FR} are kept frozen, and $DIP(z)$ is the differentiable image parametrization constraint.

Differentiable Image Parametrizations (Mordvintsev et al., 2018). When optimizing an input image that maximizes some activation of a neural net, it was shown that further parametrizing the image e.g. by optimizing its complex spectrum as opposed to RGB pixel values, can improve the interpretability of the generated images.

Deep image prior (Ulyanov et al., 2018). Related to the above, the authors show that randomly-initialized convolutional neural nets with random noise inputs can serve as a strong prior towards natural images. To that end, we use the proposed Deep Image Prior as our image parametrization. Here, the input image is generated as $I = m_{UNET}(z)$, where m_{UNET} is a U-Net (Ronneberger et al., 2015) CNN, z is a tensor of random white noise, and the optimization of eq. (1) is only over the parameters of the U-Net.

Gated sparse auto-encoder (Rajamanoharan et al., 2024). The aim of sparse autoencoders is to decompose highly compressed templates, in ideally case resulting sparse features, which encode information in a more disentangled, interpretable manner. To obtain sparse template representations, we implement gated SAE as it was shown to improve dictionary learning. The idea of G-SAE in comparison to vanilla SAE is to replace the simple ReLU encoder with a gated ReLU encoder by separating the gating and magnitude encoding roles,

$$\tilde{\mathbf{f}}(\mathbf{x}) := \underbrace{\mathbb{1}\left[(\mathbf{W}_{\text{gate}}(\mathbf{x}-\mathbf{b}_{\text{dec}})+\mathbf{b}_{\text{gate}})>0\right]}_{\mathbf{f}_{\text{gate}}(\mathbf{x})} \odot \underbrace{\text{ReLU}(\mathbf{W}_{\text{mag}}(\mathbf{x}-\mathbf{b}_{\text{dec}})+\mathbf{b}_{\text{mag}})}_{\mathbf{f}_{\text{mag}}(\mathbf{x})} \tag{2}$$

where $\mathbb{1}$ stands for the Heaviside step function, and $\mathbf{W}_{\text{gate}}, \mathbf{W}_{\text{mag}}, \mathbf{b}_{\text{dec}}$ and $\mathbf{b}_{\text{mag}}$ are trainable parameters. The decoder is again a single linear layer.

We design G-SAE such that it expands templates $x \in \mathbb{R}^{512}$ by a factor of 128, producing sparse encodings of $f(x) \in \mathbb{R}^{65,536}$. Since the step function is non-differentiable, but the gate and magnitude weights are tied, after training the G-SAE, we optimize I to maximize $\mathbf{W}_{\text{gate}}(\mathbf{x}-\mathbf{b}_{\text{dec}})$ in practice.

4. Experimental setup

Face recognition models and templates. State-of-the-art face recognition models are typically trained on pre-aligned face images, and are known to be sensitive to misalignment. To extract templates x, we align face images from the VGGFace2 dataset using the MTCNN (Zhang et al., 2016) keypoint detector and down-scale them to 112×112 pixels. We then extract the face templates of all the images in advance. This speeds up the autoencoder training considerably, as it obviates the need to run the face recognition model at the same time.

Gated sparse auto-encoder (G-SAE). Using the extracted templates, we train the autoencoder using the loss function proposed by (Rajamanoharan et al., 2024), with sparsity factor $\lambda = 10^{-7}$. We use a batch size of 1024, and train the autoencoder using the AdamW (Loshchilov & Hutter, 2017) update rule with an initial learning rate of 5×10^{-4} and weight decay factor of 10^{-2}. The training run lasts 100 epochs, during which we decay the learning rate towards 10^{-6} using the cosine schedule.

Input image optimization. To optimize input images, which maximized the activation of a targeted feature in the sparse encoding, we use AdamW optimizer with a learning rate of 10^{-4}, and we run the optimization for $4,096$ steps. We experimentally determine a batch size of 1 results in the most interpretable images, since running the optimization with larger batch sizes results in very similar images. To produce multiple images of the same sparse encoding component, we re-run the optimization with different random parameter initialization of the DIP instead.

5. Experiments and Results

Evaluating the sparse feature composition. We first evaluate how many of the images I activate each of the sparse features $f(x)_i$. The results are presented in the Figure 2. We note that unlike previous results from language models (Bricken et al., 2023), the features are mainly divided into two clusters: a small cluster of $451/65,536$ features that are activated by $40\% - 48\%$ of the dataset images at non-zero magnitudes, and a much larger cluster of very low-density features that are only activated by $1 - 20$ out of the 3.2×10^6 dataset images. We investigate the difference between the features of those two clusters in subsequent experiments.

Reconstruction and redundancy. Next, we evaluate the redundancy of the sparse representation. We note that only 39.64% of the sparse features are left alive (i.e., activated by at least one dataset image) by the end of the training. This is despite using the neuron re-sampling procedure proposed by (Bricken et al., 2023). The mean L_0 norm of $f(x)$ is 105.2. We also evaluate performance of the trained G-SAE

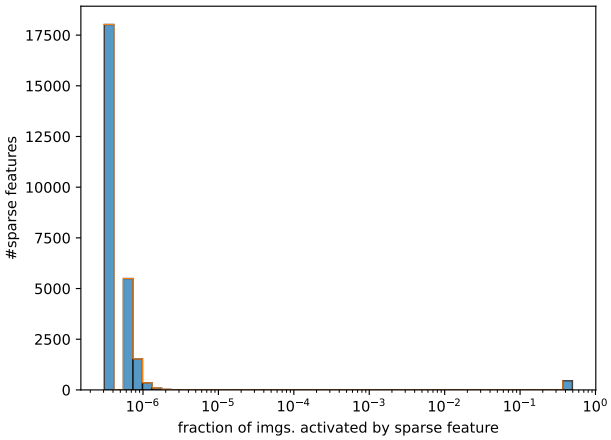

Figure 2. Feature density histogram. Note the logarithmic x-axis.

in terms of the reconstruction loss. The average MSE over the dataset is $\mathbb{E}_x||x - \hat{x}||_2^2 = 7.13 \times 10^{-6}$, whereas we have pre-scaled the face templates x such that $\mathbb{E}_x||x||_2^2 = 1$ by applying a uniform scaling factor to the entire dataset.

Loss explained. We also evaluate the quality of the reconstructed templates for downstream tasks (i.e., face verification). We attempt to decode the templates using either only the high-density feature cluster (i.e., the rightmost features in Figure 2), or the entire sparse representation. Using the reconstructed features, we perform the standard LFW (Huang et al., 2007) face verification experiment. LFW is known as a "solved dataset" in the sense that modern face recognition models approach 100% performance. However, it fits our purpose here as we are primarily interested in the amount of performance degradation introduced by the autoencoder given different settings. We present the results in table 1. We note that the features from the high density cluster account for most of the loss explanation.

Table 1. LFW results using decoded templates.

Setting	**Verification accuracy** $(\mu \pm \sigma)$
Original templates $x = m_{FR}(I)$	0.9948 ± 0.0017
Decoded from high-density cluster of $f(x)_i$	0.9128 ± 0.0110
Decoded from low-density cluster of $f(x)_i$	0.6120 ± 0.0511
Decoded from entire $f(x)$	0.9655 ± 0.0062

Interpreting individual sparse features. In this set of experiments, we analyse individual components of the sparse encoding $f(x)_i$ by optimising input images to activate them, as well as comparing with the dataset images that produce highest magnitude activations. We find that many of the features in the high density cluster correspond to general face attributes, and some are also explainable using the op-

Table 2. Qualitative feature interpretation.

$i,\ f(x)_i$	Comment	Quality-equivariant	Explanation from DIP
2396	Young east Asian men	✓	✓
2500	Middle-aged men with rectangular glasses	✗	✓
7715	Young east Asian women with long hair	✓	✗
9341	Men with occluded or poorly-visible eyes	✗	✗
12230	Dionne Warwick, age-invariant	✓	✗
28105	Women with occluded or closed eyes	✗	✓
28177	Older white women	✗	✗
28617	East Asian women with short or tied hair	✓	✗
48971	"Bangs" hairstyle	✗	✗
55143	White women with pronounced nasolabial folds	✓	✓
58513	Bianca Balti	✓	✗
60871	Women or men with thick dark eyebrows	✗	✓
63922	Men with a V-shaped hairline	✗	✗

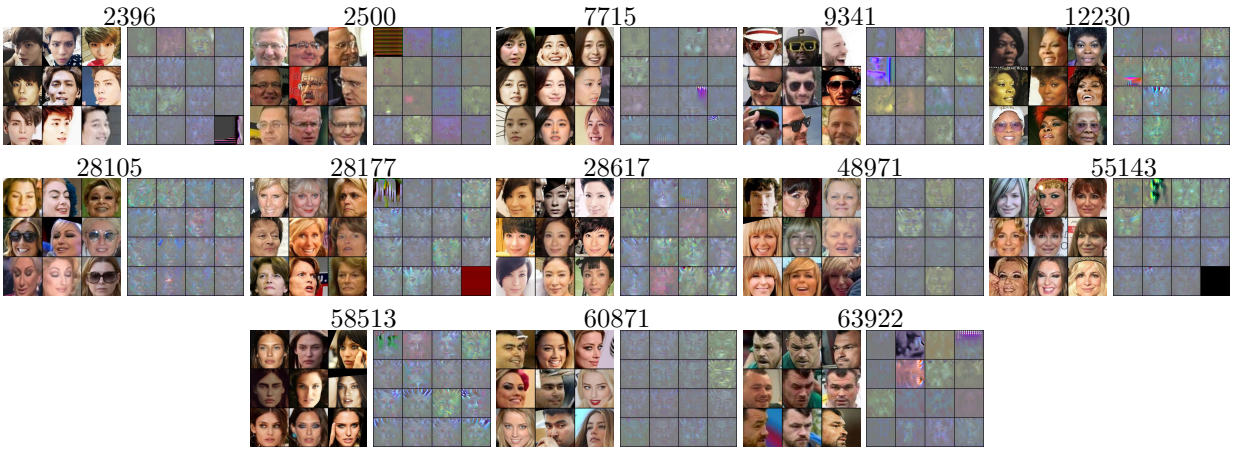

Figure 3. Dataset samples and optimised images for selected feature components. Best viewed zoomed in.

timized images. We note that most of these "face attribute neurons" are pose-invariant, activating with similar magnitudes given images of different yaw poses. We also note that some are quality-equivariant, i.e., they produce higher activation magnitude for sharper, high-quality images, whereas others are quality invariant in the sense that there is no correlation between perceptual image quality and feature activation magnitude. We also examine whether the interpretation derived from the qualitative inspection of dataset images that activate a given feature is apparent from the optimised input images. We summarise the properties of some identified features in table 2 and present visualizations in Figure 3. We note that within the high density feature cluster, almost none of the features correspond to a specific identity. To the contrary, all of the low density cluster features examined correspond to a specific person and are less readily interpretable.

6. Conclusion

We have demonstrated the successful application of LLM mechanistic interpretability techniques to face recognition models. By converting face image templates into sparse representations with gated sparse autoencoders, we have achieved a more interpretable form of the encoded information.

We have shown the encoded representations consist of two feature clusters - a small group of high-density features that encode various general face attributes, and a large amount of low-density features that probably encode information specific to individual identities from the training set.

In some cases, the qualitative feature interpretation of the dataset samples can be confirmed by generating input images using differentiable image parametrizations and the deep image prior.

Impact Statement

This paper presents work aimed at advancing the interpretability of face recognition models. Research in this direction can contribute to more trustworthy and accountable applications in biometric identification. The ethical aspects of our work include promoting fairness and reducing bias in automated decision-making, as well as providing clearer insights into model operations, which is crucial for high-stakes applications such as law enforcement.

While there are many potential societal consequences of our work, including improved user trust and regulatory compliance, we believe that the positive implications of making face recognition models more interpretable and transparent should be highlighted. This increased transparency can help mitigate issues related to misuse and bias, ensuring these technologies are deployed ethically and responsibly.

Acknowledgement

Research presented here was funded by the Slovenian Research and Innovation Agency (ARIS) research Programme P2-0250 "Metrology and biometric systems", and the ARIS research project J2-50069 "Mechanistic Interpretability for Explainable Biometric AI (MIXBAI)"

References

Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, pp. 2, 2023.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

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

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mordvintsev, A., Pezzotti, N., Schubert, L., and Olah, C. Differentiable image parameterizations. *Distill*, 2018. doi: 10.23915/distill.00012. https://distill.pub/2018/differentiable-parameterizations.

Palit, V., Pandey, R., Arora, A., and Liang, P. P. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2856–2861, October 2023.

Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., and Deng, W. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024.

Rajpal, A., Sehra, K., Bagri, R., and Sikka, P. Xai-fr: explainable ai-based face recognition using deep neural networks. *Wireless Personal Communications*, 129(1): 663–680, 2023.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Williford, J. R., May, B. B., and Byrne, J. Explainable face recognition. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, pp. 248–263, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58621-8.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503, 2016.