Isolating Racial Semantics in a Vision Transformer Using Semi-Supervised Sparse Autoencoders

Ozan Bayiz
ozanbayiz@berkeley.edu

Kapil Malladi
kapilmalladi@berkeley.edu

Charlie Cooper
charlie.c@berkeley.edu

Raiyan Hammad Ausaf
raiyanausaf14@berkeley.edu

Abstract

We investigate how racial semantics are encoded in the latent representations of a Vision Transformer (ViT), extending recent work on sparse autoencoder (SAE) analysis of large language models (LLM) by Cunningham et al. [1] and the GG Claude project [2]. Prior research has used unsupervised SAEs to retrospectively uncover human-interpretable features through various methods (CORRESPONDING SECTION OF GGC)[5], including OpenAI's AutoInterpretability scores for LLMs (Bills, S., et al.)[4]. These scores are ill-suited for our vision task. We diverge from the two previous research efforts by specifying our semantic feature of interest—race—*before* using an SAE to learn the SDFs. We begin by applying an SAE to final-layer ViT activations and visualizing the SDFs using dimensionality reduction. Due to insufficient separation by race, we pursue two approaches: (1) training a semi-supervised SAE with an auxiliary classification loss for race, and (2) modeling class-wise latent distributions and perturbing representations along inter-class "race vectors." These perturbations are used to test causal influence on predictions, which we further validate with activation patching. Our findings offer a pathway toward more targeted interpretability methods in vision models, with implications for fairness and representation auditing.

1 Introduction

Understanding how sensitive attributes like race are encoded in neural network representations is central to ongoing discussions about fairness, transparency, and interpretability in machine learning. While prior work in large language models (LLMs) has used Sparse Autoencoders (SAEs) to uncover human-interpretable latent features, applications to vision models remain underexplored. This project investigates whether race-related features are encoded in the latent space of a Vision Transformer (ViT), and whether these features can be isolated using either architectural modifications or systematic post hoc analysis.

2 Related Work

This work builds on several lines of research. Cunningham et al. [1] and the GG Claude project [2] apply SAEs to LLM activations, interpreting learned dictionary features retrospectively. These efforts are motivated by the superposition hypothesis [3], which argues that human-interpretable concepts are entangled across features in latent space. Recent works have explored ways to disentangle these through post hoc probes and feature attribution [4].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

However, most of these methods are fully unsupervised and do not target a specific semantic concept in advance. Our project diverges by taking a semi-supervised route, explicitly focusing on racial attributes in ViT embeddings and seeking to isolate dictionary features that encode these semantics.

3 Problem Statement

We aim to determine whether racial information is embedded in the latent space of a ViT and, if so, whether it can be disentangled and manipulated. Specifically, we ask:

1. Can semi-supervised SAEs meaningfully change the sparse dictionary features (SDFs) to be sensitive to a predetermined concept?
2. Can we construct interpretable directions in the SAE latent space that correspond to transitions between racial categories?
3. Do these directions causally affect race classification predictions when applied as perturbations?

4 Approach

We first trained a standard Sparse Autoencoder (SAE) on the output of the final layer of the visual encoder block of a ViT. These representations were projected into 2D using PCA, UMAP, and t-SNE to assess clustering by race.

INSERT IMAGES OF PCA, t-SNE & UMAP

Initial results showed little structure when colored by race label, suggesting that unsupervised SDFs may not align with our feature of interest. We then pursued two strategies:

4.1 Semi-Supervised SAE

We defined a new loss function:

$$L(x, y) = \text{reconstruction loss}(x, \hat{x}) + \alpha \cdot \text{label loss}(\text{classify}(\text{enc}(x)), y)$$

Where x is the vision encoder output activation and y is the race label. This encourages the encoder to generate compressed representations predictive of racial class labels while preserving reconstruction quality.

4.2 Alternative Analysis of SDFs

We propose a new statistical procedure:

- Model each race's encoded representations as multivariate Gaussians.
- If clusters aren't sufficiently disparate, we identify low-variance dimensions within each class to select stable, race-distinctive features by zeroing out high-variance entries.
- Define "race vectors" as directions between class means in this subspace, including from/to a "neutral race" (centroid of means).
- Perturb encoded activations along these race vectors and evaluate changes in classifier prediction and confidence.

5 Preliminary Results and Findings

Dimensionality reduction visualizations (PCA, UMAP, t-SNE) show limited race-based clustering in the raw SAE outputs.

Variance-based filtering reveals promising feature subsets where some SDFs show low intra-class variance and high inter-class separation.

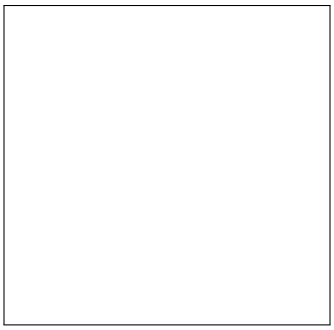

Figure 1: Sample figure caption.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Next steps include computing race vector perturbations and visualizing their effect on classification outcomes.

We also plan to use activation patching to validate whether modifying these directions causally influences downstream race classification behavior.

5.1 Tables

References

[1] Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. EleutherAI, MATS, Bristol AI Safety Centre, Apollo Research. Retrieved from `https://transformer-circuits.pub/2023/monosemantic-features/index.html`

[2] GG Claude Project. (2024). [Title TBD or leave blank]. [Institution, if known].

[3] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Chen, A., ... & Amodei, D. (2022). A Mathematical Framework for Transformer Circuits. Transformer Circuits. Retrieved from `https://transformer-circuits.pub/2022/toy_model/index.html`

[4] Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023, May 9). Language models can explain neurons in language models. OpenAI. Retrieved from `https://www.openai.com/research/language-models-can-explain-neurons`

[5] [Corresponding section of GG Claude project analysis]. [Include author(s) or institution if known, or leave as placeholder for now.]