INVESTIGATING RACIAL SEMANTICS IN VISION TRANSFORMER REPRESENTATIONS USING FAIRFACE

Ozan Bayiz
ozanbayiz@berkeley.edu

Kapil Malladi
kapilmalladi@berkeley.edu

Charlie Cooper
charlie.c@berkeley.edu

Raiyan Hammad Ausaf
raiyanausaf14@berkeley.edu

April 17, 2025

ABSTRACT

We need more GPUs.

1 "Coherent Story" Justifying our "Systematic Investigation"

Recent high-profile failures of commercial vision systems — most notably the infamous Google Photos misclassification of Black individuals as "gorillas" — underscore the need for racial interpretability in deep learning models. Despite improvements in classification accuracy, modern vision systems often lack transparency in how sensitive attributes like race are internally represented. Our project aims to identify these ideas within a ViT model, and discover key features used by the model for classifying race. We use modern methods of mechanistic interpretability, such as sparse autoencoders and the superposition hypotheses to come up with, and empirically back, hypotheses.

2 Method

2.1 Linear Probe

2.2 SAE

2.3 Collect Stats

Table and visualization of the means in latent vector space

2.4 Intervene

What becomes an interesting question is how moving in the latent space truly affects the racial classification. It seems entirely plausible that moving along a vector defined between white and asian could bring the classification closer to Black. Or the average activation (of all racial means) being actually classified as Black or Indian, etc. due to proximity to the mean. If I am understanding this correctly, we will still be able to do a sanity check based on classifying the mean activations of each race (I really hope that works)

3 Components

3.1 Florence-2

3.2 SAE

3.3 Fairface