The Missing Curve Detectors of InceptionV1: Applying Sparse Autoencoders to InceptionV1 Early Vision

Liv Gorton

TL;DR: We demonstrate sparse autoencoders discover new, interpretable features that aren't recovered when studying the neurons.

Introduction

- The original project of mechanistic interpretability was to understand the convNet, InceptionV1.
- However, progress was limited by polysemantic neurons that respond to multiple unrelated stimuli.
- It has been hypothesised that polysemanticity is due to superposition, where features are represented by combinations of neurons.
- Over the last year, progress has been made on extracting features in superposition from language models. What if we brought that work back to InceptionV1?

Methods

We trained sparse autoencoders on largely following the methodology of Connerly et al.

Decompose activation vector x into:

$$x \simeq \sum_i \underbrace{b}_{\text{decoder bias}} + \underbrace{f_i(x)}_{\text{feature activation}} \underbrace{d_i}_{\text{feature direction}}$$

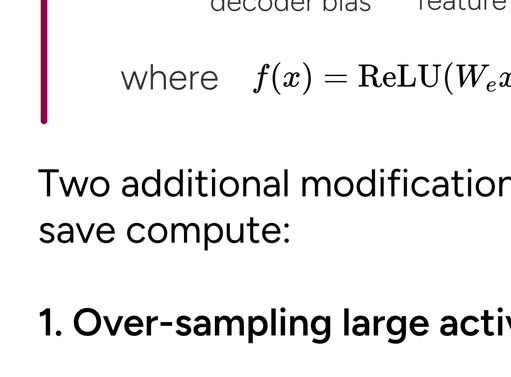
where $f(x) = \text{ReLU}(W_e x + b_e)$

Two additional modifications were made to save compute:

1. Over-sampling large activations

- The majority of image positions produce small activations (e.g., backgrounds).
- These small activations can be thought of as less relevant to what the layer is representing.
- We sample activations from each image proportional to their activation magnitude to avoid compute spend modelling these small activations.

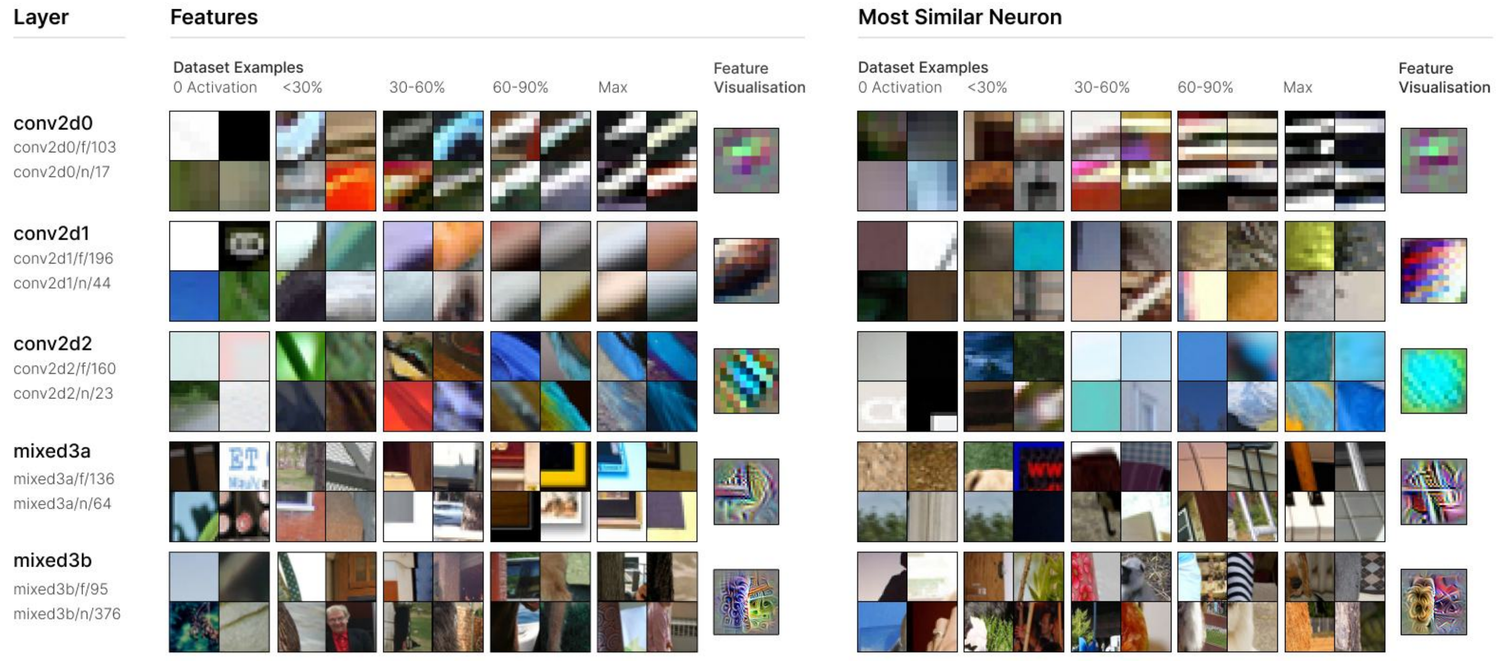
2. Branch specific SAEs



- SAEs for mixed3a and mixed3b were trained on the **3x3** and **5x5** branches of InceptionV1.
- Motivated by intuition that features are likely isolated to the branch with the most advantageous filter size.

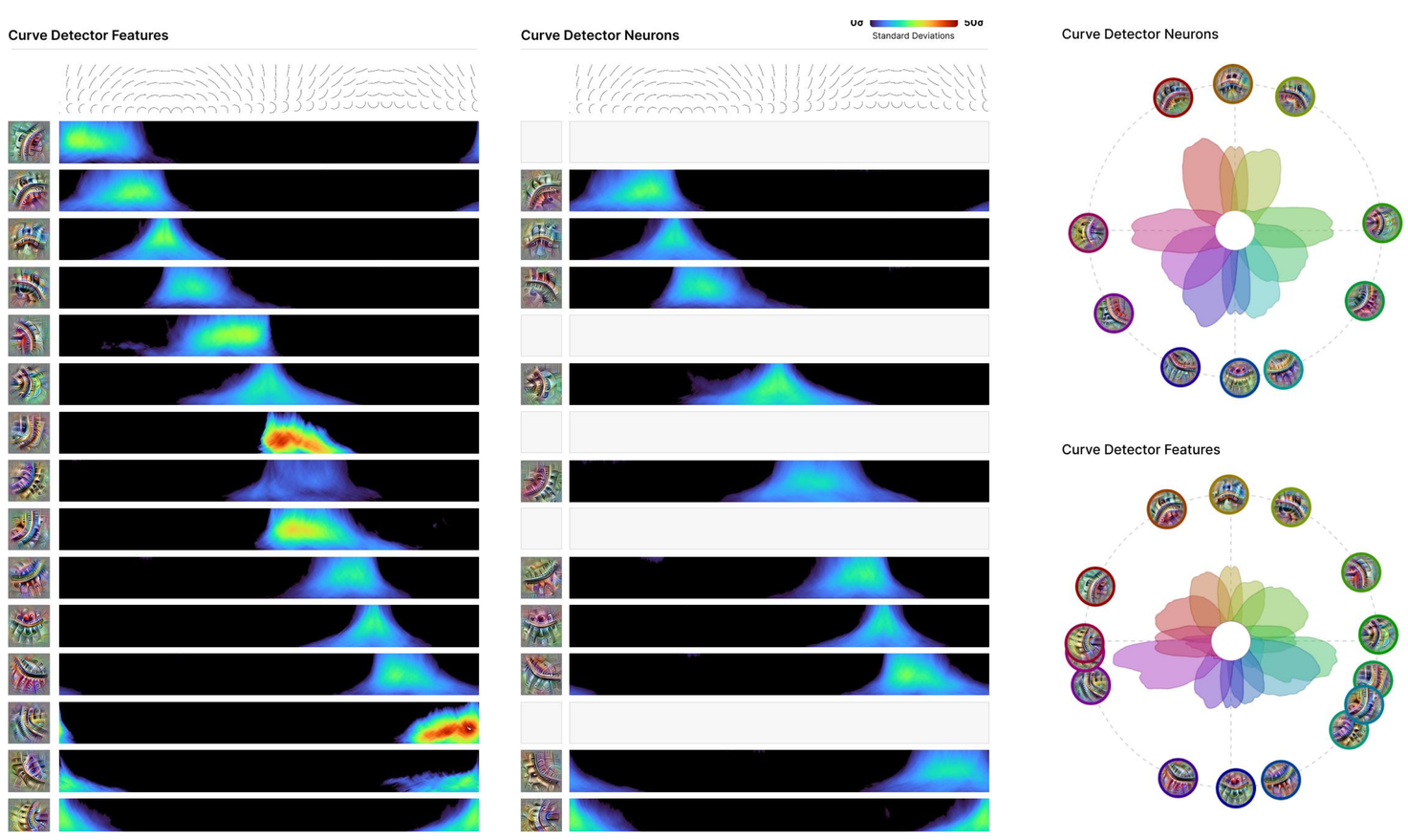
Results

1) SAEs Learn New, Interpretable Features



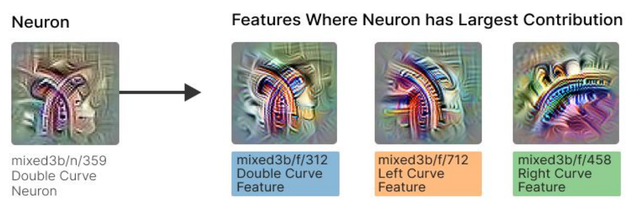
For each "early vision" layer in InceptionV1, features are discovered which are interpretable features across the entire activation spectrum.

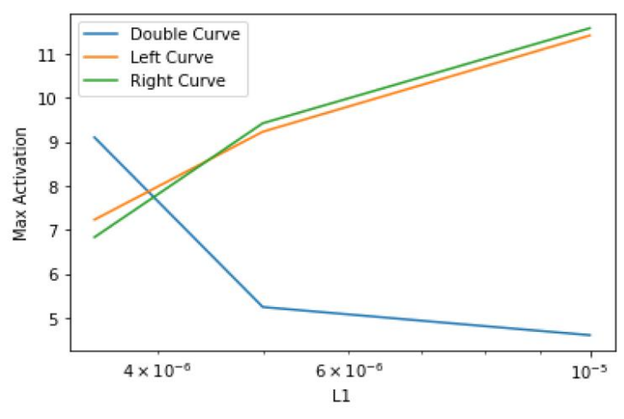
2) SAEs Discover Additional Curve Detectors



Curve detector features were found for previously unidentified orientations that fill in the gaps that occur when just studying the neurons.

3) SAEs Split Polysemantic Neurons





- Double curve detectors, previously believed to be polysemantic, decompose into three features: two monosemantic curve detectors and a double curve detector.
- The double curve detector shrinks as L1 coefficient increases.

PAPER