arXiv:2411.08166v1 [cs.LG] 12 Nov 2024

Tackling Polysemanticity with Neuron Embeddings

Alex Foote [1 2]

Abstract

We present neuron embeddings, a representation that can be used to tackle polysemanticity by identifying the distinct semantic behaviours in a neuron's characteristic dataset examples, making downstream manual or automatic interpretation much easier. We apply our method to GPT2-small, and provide a UI for exploring the results. Neuron embeddings are computed using a model's internal representations and weights, making them domain and architecture agnostic and removing the risk of introducing external structure which may not reflect a model's actual computation. We describe how neuron embeddings can be used to measure neuron polysemanticity, which could be applied to better evaluate the efficacy of Sparse Auto-Encoders (SAEs).

1. Introduction

Mechanistic Interpretability (MI) aims to decompose neural networks into their constituent parts and understand how these parts interact to create the behaviour of the network, with the ultimate goal of understanding models in enough detail to determine whether they're safe to deploy. One of the key suppositions of MI is that it's possible to break models apart into meaningful units, often called features (Olah et al., 2020). One natural basis for these units is the neuron, and visualisation techniques developed for vision models had significant success in understanding the function of many neurons, as well as how they compose to implement increasingly complex behaviours (Cammarata et al., 2020).

However, a major obstacle to this approach is the fact that neurons often respond to several completely distinct concepts, a phenomenon called polysemanticity. This makes it much harder to find a clean and simple explanation for a neuron's behaviour, and undermines the idea that neurons are the natural basis for decomposing a model. Polysemanticity is particularly prevalent in language models, and has made interpreting their MLP layers a significant challenge (Elhage et al., 2022a).

One common method for interpreting the behaviour of a neuron in a language model is to collect and study the dataset examples which cause the highest neuron activation. Patterns in a neuron's dataset examples provide an indication of what the neuron responds to. However, polysemanticity makes these dataset examples much harder to interpret, as there are often many separate behaviours to understand, some of which may be related and others entirely distinct. This becomes increasingly challenging as you collect examples further down the activation spectrum, which is important for gaining a complete understanding of a neuron, but often reveals a wider range of behaviours (Bolukbasi et al., 2021).

To tackle the problem of polysemanticity, we introduce **neuron embeddings**, which capture the information that a given neuron is responding to in a given input. Given a neuron which we're trying to understand and an input which causes that neuron to activate, we define the neuron embedding of the input as the element-wise product of the vector representation that the neuron receives and the neuron's input weights. We show that this representation can be used to cluster a neuron's dataset examples, making it possible to disentangle the neuron's behaviour into it's constituent parts. Dataset examples are used for both manual and automated interpretability (Bills et al., 2023; Foote et al., 2023), so making them easier to interpret has significant potential benefit for a variety of downstream applications. We apply this method to GPT2-small (Radford et al., 2019) and provide case studies on individual neurons, as well a website for exploring the results for the full model [1].

Crucially, neuron embeddings also allow us to measure a proxy for a neuron's degree of polysemanticity by computing simple metrics on the geometry of the points and the clusters that form. Sparse Auto-Encoders (SAEs) (Bricken et al., 2023) are a promising technique for dealing with polysemanticity, which learn to disentangle a layer of neurons into a wider, sparse MLP layer with monosemantic neurons. However, we lack effective metrics for evaluating the quality of SAEs, instead relying on simple heuristics like reconstruction error and activation sparsity, as well as time consuming manual analysis. Neuron embeddings may be

[1]Ripjar [2]Apart Research. Correspondence to: Alex Foote <alexjfoote@icloud.com>.

Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] https://feature-clusters.streamlit.app