Steering Vectors

Julian Bernado

March 4, 2025

1 Introduction

The goal of this document is to be a self-contained theoretical briefer of steering vectors along with the DiffMean/CAA and ReFT-r1 methods of producing steering vectors.

2 Notation

Let $T : \mathbb{R}^{C\times d} \to \mathbb{R}^{C\times V}$ be a transformer composed of L transformer blocks with vocabulary $\mathcal{V}$ such that $|\mathcal{V}| = V$ and context-length C. T maps C tokens with embedding dimension d

$$h^{(0)} \in \mathbb{R}^{C\times d} = \begin{bmatrix} h_{11}^{(0)} & ... & h_{1d}^{(0)} \\ ... & ... & ... \\ h_{C1}^{(0)} & ... & h_{Cd}^{(0)} \end{bmatrix}$$

into C logits with vocabulary size V

$$h^{(L+1)} \in \mathbb{R}^{C\times V} = \begin{bmatrix} h_{11}^{(L+1)} & ... & h_{1V}^{(L+1)} \\ ... & ... & ... \\ h_{C1}^{(L+1)} & ... & h_{CV}^{(L+1)} \end{bmatrix}.$$

For $\ell = 1, ..., L$, we represent the output of the ℓth transformer block T_ℓ as $h^{(\ell)} \in \mathbb{R}^{C\times d}$ with residual stream as

$$h^{(\ell)} = h^{(\ell-1)} + T_\ell\left(h^{(\ell-1)}\right)$$

with output matrix $W \in \mathbb{R}^{V\times d}$ such that $h^{(L+1)} = h^{(L)}W^\top$ mapping to the logits. Then, we have that the whole operation is

$$T\left(h^{(0)}\right) = \left[h^{(0)} + \sum_{i=1}^{L} T_i(h_{i-1})\right] W^\top$$

For notational utility later on, let $A_\ell(\cdot, \cdot)$ map a transformer and its input to its activations on the Cth token after the ℓth layer:

$$A_\ell(T, h_0) = \left[h^{(\ell)}\right]_{C,\cdot}^\top \in \mathbb{R}^d$$

3 Steering Vectors

A steering vector $v \in \mathbb{R}^d$ along with steering coefficient $\alpha \in [-1, 1]$ can be added to the residual stream at layer ℓ at token position $c \in 1, ..., C$. Typically, we add the steering vector to the final token position C. This produces the alternate intervened-upon $h_*^{(\ell)}$ defined as

$$h_*^{(\ell)} = h^{(\ell-1)} + T_\ell\left(h^{(\ell-1)}\right) + \alpha \begin{bmatrix} \mathbf{0} \\ v^\top \end{bmatrix}$$

Passing this intervened state onto future layers, we define intervened transformer T_* with

$$h_*^{(\ell+k)} = h_*^{(\ell+k-1)} + T_\ell\left(h_*^{(\ell+k-1)}\right)$$

for $k = 1, ..., L - \ell$ and $h_*^{(L+1)} = h_*^{(L)} W^\top$. It's worth noting here that when applying the steering vector, we do so at a single layer and let the effects propagate throughout the later layers rather than applying the steering vector at multiple layers.

The goal of a proper steering vector is to represent some meaningful feature of the model's possible response. Work has been done to extract steering vectors corresponding to a direction of truthfulness [MT24], but potential directions for steering extend far beyond this. For a given concept targeted by steering, adding a steering vector should causally increase the likelihood of outputting a response with the desired behavior. You may also generally target higher performance on some scale corresponding to presence of the concept. The next two sections will explore different methods for producing such a steering vector.

4 DiffMean / Contrastive Addition Activation

The method is first presented in [Pan+24]. We'll discuss the steps to calculate the steering vector then discuss the specifics of forming a steering dataset.

4.1 Learning the DiffMean Vector

Let $\mathcal{D} = (p_i^+, p_i^-)_{i=1}^n$ be a set of n prompts representing positive and negative examples of some targeted concept. In particular, $p_i^+, p_i^- \in \mathbb{R}^{C \times d}$ are tokenized contexts of the same length. Furthermore, the first $C - 1$ rows of p_i^+ and p_i^- will be identical; the two only differ in final token. Now, fix $\ell \in \{1, ..., L\}$. Then, define v_ℓ as

$$v_\ell = \frac{1}{n} \sum_{i=1}^{n} A_\ell(T, p_i^+) - A_\ell(T, p_i^-)$$

Then, we apply v_ℓ at layer ℓ as described in the previous section to net the DiffMean-steered transformer T_{DiffMean}. Generally, we can think of steering methods as a function from an existing transformer and dataset $\mathcal{D}$ to a new steered transformer.

For the choice of layer and α, a grid (or other) search can be performed to select a layer ℓ^* and steering magnitude α^* that optimally produces the desired behavior.

4.2 Creating a Steering Dataset

To construct the dataset $\mathcal{D}$, the authors of [Pan+24] apply prompts of the form:

p_i^+ : "[INST] Neutral prompt.
Choices:
(A) Positive example of behavior.
(B) Negative example of behavior.
Choice: (A"

p_i^- : "[INST] Neutral prompt.
Choices:
(A) Positive example of behavior.
(B) Negative example of behavior.
Choice: (B"

Here we see how the two prompts can represent positive and negative examples of the same behavior while remaining the exactly same number of tokens.

5 ReFT-r1

References

[MT24] Samuel Marks and Max Tegmark. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets". In: (2024). arXiv: 2310.06824 [cs.AI]. URL: https://arxiv.org/abs/2310.06824.

[Pan+24] Nina Panickssery et al. "Steering Llama 2 via Contrastive Activation Addition". In: (2024). arXiv: 2312.06681 [cs.CL]. URL: https://arxiv.org/abs/2312.06681.