Multi-Layer Activation Steering for Image Generation

AML – Midterm Presentation

01/12/2025

SAPIENZA
UNIVERSITÀ DI ROMA

Group: **STITCH**

- Davide Perniconi: 1889270
- Olja Corovencova: 2249558
- Daniele Marretta: 1985747
- Leonardo Lavezzari: 1984079

Task and Motivation

Motivation

Generative models often exhibit misaligned or uncontrollable behaviors (e.g., toxicity, bias) that are expensive to fix via retraining.

We need dynamic, precise, and low-cost control over high-level concepts encoded in a model's internal activations.

With Activation Steering we can effectively guide the model's internal computation without altering its weights with minimal computation overhead.

Task Statement

Study how an image generation model can be controlled through direct interventions on their internal activations:

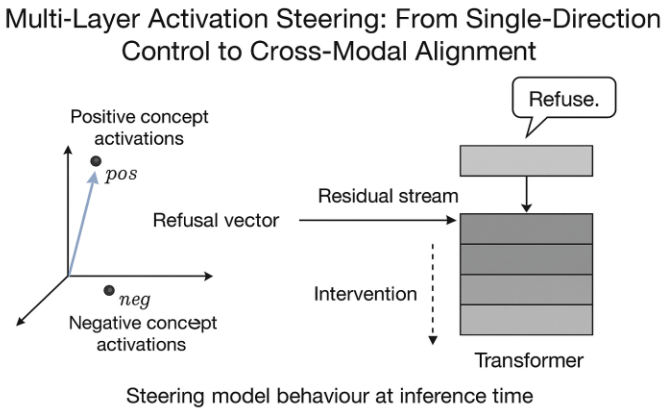
1. Apply **single** and **multi-layer steering** to Stable Diffusion 2.

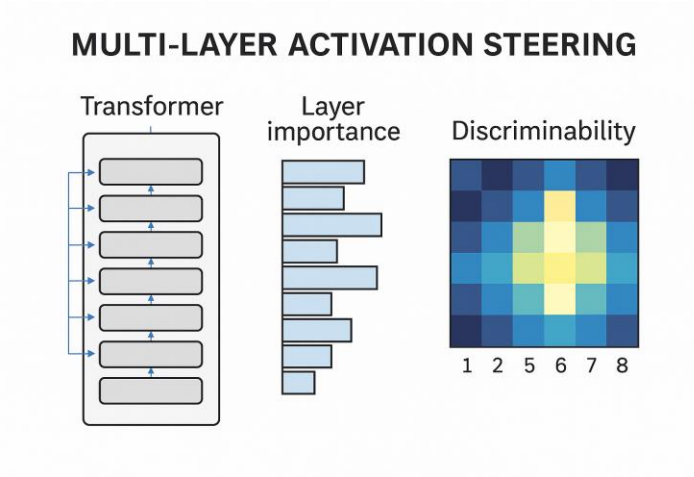
2. Experiment with new vector generation and layer choosing strategies.

Task and Motivation

Related Work

- **Refusal in Language Models Is Mediated by a Single Direction (2024**), finds refusal vector, erasing it "jailbreaks" the model, while adding it induces refusal on benign prompts.
- **LayerNavigator: Finding Promising Intervention Layers for Activation Steering (2025),** proposes a method to score each layer's "steerability" based on discriminability consistency.
- **Controlling Language and Diffusion Models by Transporting Activations (2025),** finds a minimal-cost map to move the source activation distribution to the target distribution





Methodology

Method

- Create “positive” and “negative” examples for the target behaviour.
- Steering Vector Extraction (Mean Difference(+PCA), alternatives) using the contrastive prompts.
- Applying the Steering Vector to Stable Diffusion 2.

Investigation

- Compare how single-layer and multi-layer steering would affect the model's behaviour.
- Analyse how steerability would vary in the initial, middle, and final layers.
- Explore whether LayerNavigator metrics could be applied to image generation.

Analysis

The objective will be to systematically compare the behaviour of the model before and after steering, evaluating the effect of interventions carried out at different points in the network.

To evaluate the effectiveness of steering we'll measure the model's response to negative prompts drawn from chosen datasets (I2P, CPDM)

The evaluation could include metrics such as:

1. **FID** (Fréchet Inception Distance), to ensure that steering does not degrade overall quality.
2. **GPT** evaluation (percentage on each image), to estimate how much the model would manifest the expected behaviour.
3. **CLIP Score**, to measure semantic similarity between text prompt and generated image.