Prisma

Multimodal Mechanistic Interpretability Library

What is Prisma?

Prisma is a multimodal mechanistic interpretability library currently based on TransformerLens. Prisma currently supports vanilla vision transformers (ViTs) and their vision-text counterparts CLIP.

The goal of Prisma is to make research in mechanistic interpretability for multimodal models both easy and fun. We are also building a strong and collaborative open source research community around Prisma.

While language mechanistic interpretability already has strong conceptual foundations, many research papers, and a thriving research community, non-language modalities lag behind. Given that multimodal capabilities will be part of AGI, field-building in mechanistic interpretability for non-language modalities is crucial for safety and alignment.

See this link for the Prisma GitHub repo, which contains links to onboarding tutorials.

Goals

1. Build shared infrastructure (Prisma) to make it easy to run standard language mechanistic interpretability techniques on non-language modalities.
2. Build a shared conceptual foundation for multimodal mechanistic interpretability.
3. Shape and execute on research agenda for multimodal mechanistic interpretability.
4. Build an amazing multimodal mechanistic interpretability community.
    a. Set cultural norms of this community to be highly collaborative, curious, inventive, friendly, respectful, prolific, and safety/alignment-conscious.
    b. Encourage sharing of early/scrappy research results on Discord/Less Wrong.
    c. Co-create a web of high-quality research.

Signs that our efforts are working

There is a renaissance in multimodal mechanistic interpretability research. This may take form in a cascade of related papers, blog posts, repos, and other projects. The Discord is lively. Niche sub-communities emerge.

Prisma Themes

The following are the main themes we focus on for Prisma. For the most recent specific tasks, check out our GitHub Issues.

Circuit-Level Analysis for Vision/Multimodal Infra

High-level Goals

- Build infra that makes it very easy to transfer standard mech interp techniques on vision and vision-text
    - Standard techniques include, but are not limited to: activation patching, direct logit attribution, logit lens, and attention head visualization.
- Transfer language mech interp to the peculiarities of vision and vision-text

Tasks

- Adapt new models to the repo
    - Vision Transformer Variants
        - [ ] TinyCLIP: https://github.com/soniajoseph/ViT-Prisma/issues/80
        - [ ] CLIP: https://github.com/soniajoseph/ViT-Prisma/issues/81
            - [ ] CLIP small
            - [ ] CLIP medium
            - [ ] CLIP large
        - [X] Video ViT
            - [ ] Test mech interp techniques on video ViT: https://github.com/soniajoseph/ViT-Prisma/issues/82
    - Other Vision and Vision/text models
        - [ ] Flamingo: https://github.com/soniajoseph/ViT-Prisma/issues/83
    - Expanding beyond vision to other modalities (not priority for near feature, but worth thinking about and planning for)
        - [ ] Identify candidate models and modalities that are both relevant to our repo and broader safety concerns: https://github.com/soniajoseph/ViT-Prisma/issues/83
- Creating new datasets for research purposes
    - [ ] Adapt patch-level labels on ImageNet to a dataloader: https://github.com/soniajoseph/ViT-Prisma/issues/84
    - [ ] Create a simplified version of CLEVR

Code Quality / Accessibility / Documentation

High-level Goals

- Ensure that newcomers can easily onboard onto the repo
- Ensure that code is well-documented, clear, and bug-free

Tasks

- Accessibility

    - [X] Create coding tutorials for basic functionality
- Documentation
    - [ ] Add documentation to the functions that are actively used in the tutorial notebooks
    - [ ] Add documentation to all functions
    - [ ] Add automatic documentation
- Testing
    - [ ] Get unit test coverage to 100%

Research

- Make research progress on *research programme*
- Share notebooks with findings in Discord
- Make suggestions for improving repo to make your research even easier
- Contribute and frame your own problems

Tasks

For the most recent tasks, check out our GitHub Issues.

Related Projects

- TransformerLens
- SAE library
- TinyCLIP Mindreader

This document was inspired by Joseph Bloom's project management for SAE training. Thanks for your advice!