Team: Quilee Simeon & Gabriel Manso

Project Overview

Deep learning models are typically trained to transform raw data into representations optimized for specific tasks. Recently, two lines of research have inspired a deeper inquiry into the nature of these representations. The CLIP framework demonstrated the utility of aligning representations across modalities, using paired image-text data to train joint embeddings for cross-modal retrieval. Meanwhile, the Platonic Representation Hypothesis posits that performant models converge toward a shared statistical model of reality in their representation spaces, suggesting a potential universality underlying learned representations.

This project bridges these ideas by exploring whether representations from disparate **pre-trained unimodal neural networks** can be aligned into a shared multimodal latent space, inspired by the joint embedding approach of CLIP and motivated by the convergence hypothesis of Platonic Representations. The proposed framework uses **frozen unimodal encoders** (e.g., ResNet-18 for images, GPT-2 for text) with learned linear adapters to align representations across modalities. Our aim is to determine if such aligned representations better approximate those of larger, more performant models (e.g., DinoV2).

The ultimate goal is to prototype a scalable framework for aligning unimodal models into a shared multimodal latent space, with potential implications for multimodal representation learning and the mechanisms driving representation convergence.

Motivation

1. **Inspiration from CLIP**:
    - CLIP learns a joint multimodal embedding space using contrastive learning over paired image-text data. However, it requires end-to-end training of modality encoders, which limits scalability across many modalities.
    - This project adopts the contrastive learning paradigm but instead uses **frozen unimodal encoders** with **learnable linear adapters**, enabling modular scalability to additional modalities.

2. **Platonic Representation Hypothesis**:
    - Recent work suggests that performant models align their representation spaces, possibly converging toward a shared statistical model of reality. This project provides a testbed for this hypothesis by explicitly aligning modality-specific encoders and comparing the resulting representation space to that of a performant model.
3. **Scalable Multimodal Alignment**:
    - Most existing multimodal models, like CLIP, are trained jointly on paired data. In contrast, our approach proposes a framework where **frozen unimodal models** can be aligned post hoc using small-scale datasets and simple adapters, making it extensible to new modalities without retraining large encoders.

Research Questions and Hypotheses

Research Questions:

1. Can representations from unimodal models (e.g., ResNet-18 for images, GPT-2 for text) be linearly aligned into a shared multimodal latent space?
2. Does aligning representations across modalities produce embeddings that are closer to those of performant models (e.g., DinoV2)?
3. Does representational alignment provide empirical evidence supporting the Platonic Representation Hypothesis?

Hypotheses:

1. A latent space exists where unimodal representations can align via simple linear transformations enabling shared abstraction.
2. Representational alignment improves the similarity of learned representations to those of performant models, as measured by kernel alignment metrics.
3. Multimodal alignment reflects mechanisms driving representation convergence, supporting the Platonic Representation Hypothesis.

Proposed Methodology

Data and Encoders

1. **Dataset**:
    - We will use the Flickr30k dataset, which contains paired image-caption data, as a prototypical example of a multimodal dataset. Each image will be paired with its longest caption as its text modality.
2. **Unimodal Encoders**:
    - Pre-trained unimodal models will serve as feature extractors, including:
        - **ResNet-18** for the image modality, providing feature vectors in $\mathbb{R}^{d_x}$.
        - **GPT-2** for the text modality, providing feature vectors in $\mathbb{R}^{d_y}$.
    - Additional modalities (e.g., graphs or audio) could be incorporated in future work, with pre-trained encoders providing corresponding feature vectors.

Mathematical Framework

The world generates raw multimodal data:

$$\mathcal{D}_{\text{world}} = \left\{ \left(x^{(i)}, y^{(i)}, z^{(i)} \right) i = 1^N \right\}, \quad x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$$

where
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ represent different modalities (e.g., images, text, graphs), and x, y, z are specific instances of these modalities.

For each modality, we use frozen pre-trained encoders to extract representations:

$$X_{\text{enc}} : \mathcal{X} \to \mathbb{R}^{d_x}, \quad Y_{\text{enc}} : \mathcal{Y} \to \mathbb{R}^{d_y}, \quad Z_{\text{enc}} : \mathcal{Z} \to \mathbb{R}^{d_z}.$$

Learned Adapters

To align representations into a shared latent space $\mathbb{R}^{d_e}$, we introduce **linear adapters**:

$$W_x : \mathbb{R}^{d_x} \to \mathbb{R}^{d_e}, \quad W_y \quad : \mathbb{R}^{d_y} \to \mathbb{R}^{d_e}, \quad W_z : \mathbb{R}^{d_z} \to \mathbb{R}^{d_e}.$$

These adapters are the only trainable components of the system, and their

weights are optimized using a
contrastive loss.

Contrastive Learning Objective

The contrastive loss ensures that representations from the same data point across modalities are close in the shared space, while unrelated representations are farther apart:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\langle W_x X_{\text{enc}}(x^{(i)}), W_y Y_{\text{enc}}(y^{(i)}) \rangle / \tau\right)}{\sum_{j=1}^{N} \exp\left(\langle W_x X_{\text{enc}}(x^{(i)}), W_y Y_{\text{enc}}(y^{(j)}) \rangle / \tau\right)}$$

where
τ is a temperature hyperparameter.

Kernel Alignment

To evaluate the alignment of our learned multimodal representations, we define kernels for both the unimodal encoders and the shared representation space:

1. **Unimodal Kernels**:

   For each modality (e.g., $\mathcal{X}$, $\mathcal{Y}$), the kernel induced by the pre-trained encoder is defined as:

   $$K_X(i,j) = \langle X_{\text{enc}}(x^{(i)}), X_{\text{enc}}(x^{(j)}) \rangle, \quad K_Y(i,j) = \langle Y_{\text{enc}}(y^{(i)}), Y_{\text{enc}}(y^{(j)}) \rangle,$$

   where
   $x^{(i)}, x^{(j)} \in \mathcal{X}$ and $y^{(i)}, y^{(j)} \in \mathcal{Y}$.

2. **Aligned Multimodal Kernel**:

   After training the linear adapters (W_x, W_y, W_z), we compute the kernel for the aligned multimodal representations:

   $$K_{\text{repr}}(i,j) = \langle r^{(i)}, r^{(j)} \rangle, \quad r^{(i)} = W_x X_{\text{enc}}(x^{(i)}) \in \mathbb{R}^{d_e}.$$

3. **Performant Model Kernel**:

   The kernel of a larger, performant model (e.g., DinoV2) is used as a reference:

   $$K_{\text{DinoV2}}(i,j) = \langle \text{DinoV2}(x^{(i)}), \text{DinoV2}(x^{(j)}) \rangle.$$

Kernel Alignment Metric:

A kernel alignment metric, $m : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$, measures the similarity between two kernels, i.e., how similar the distance measure induced by one representation is to the distance measure induced by another.

Steps in Kernel Comparison:

1. **Before Training**:
   - Compute alignment metrics $m(K_X, K_{\text{DinoV2}})$, $m(K_Y, K_{\text{DinoV2}})$, and $m(K_Z, K_{\text{DinoV2}})$ for the unimodal kernels relative to the performant model kernel.
   - Take the average across modalities to measure initial similarity.
2. **After Training**:
   - Compute the alignment metric $m(K_{\text{repr}}, K_{\text{DinoV2}})$ for the shared representation kernel relative to the performant model kernel.

Key Hypothesis:
We hypothesize that
$m(K_{\text{repr}}, K_{\text{DinoV2}}) >$
$\text{avg}(m(K_X, K_{\text{DinoV2}}), m(K_Y, K_{\text{DinoV2}}), m(K_Z, K_{\text{DinoV2}}))$, demonstrating that aligning unimodal representations into a shared space brings them closer to the performant model's representation.

Expected Contributions

This project will contribute to the field of multimodal representation learning by:

1. **Developing a scalable framework**:
   - A modular approach to aligning pre-trained unimodal models into a shared multimodal latent space using only linear adapters.
2. **Testing representation convergence**:
   - Empirical evidence supporting or challenging the Platonic Representation Hypothesis by analyzing whether representational alignment improves kernel similarity with performant models.

3. **Providing insights into multimodal integration**:
    - Demonstrating that aligning representations across modalities can reduce modality-specific noise and increase abstraction, even with frozen unimodal encoders.

By combining the joint embedding principles of CLIP and the theoretical claims of the Platonic Representation Hypothesis, this project aims to uncover universal principles of representation learning while providing practical tools for integrating disparate unimodal models into multimodal systems.
