Aligning Pre-trained Unimodal Encoders into a Multimodal Latent Space

Quilee Simeon & Gabriel Manso

Project Overview

Deep learning models are typically trained to transform raw data into representations optimized for specific tasks. Two lines of research inspire this project. The **CLIP** framework [Radford et al., 2021] demonstrated the utility of aligning representations across modalities via joint embeddings for cross-modal retrieval. Separately, the **Platonic Representation Hypothesis** [Huh et al., 2024] suggests that performant models converge toward a shared statistical model of reality in their representation spaces, hinting at universality in learned representations.

This project bridges these ideas by aligning representations from **pre-trained unimodal encoders** (e.g., ResNet-18 for images, GPT-2 for text) into a shared multimodal latent space. Unlike CLIP, which trains modality encoders end-to-end, we freeze the encoders and learn linear adapters to align representations. We hypothesize that these aligned representations approximate those of larger, performant models (e.g., DinoV2), providing insights into the mechanisms driving representation convergence.

Motivation

1. **Inspiration from CLIP:** CLIP uses contrastive learning over image-text pairs to train joint embeddings. Our approach generalizes this to more modalities using frozen encoders with trainable adapters for modular scalability. 2. **Platonic Representation Hypothesis:** This work tests whether aligning unimodal encoders can produce representations similar to those of performant models, supporting the hypothesis of representation convergence. 3. **Scalable Multimodal Alignment:** Instead of joint training, we align frozen unimodal models post hoc using small datasets, enabling extensions to new modalities without retraining large encoders.

Research Questions and Hypotheses

Research Questions:

1. Can unimodal representations (e.g., ResNet-18 for images, GPT-2 for text) be aligned into a shared multimodal latent space?
2. Does alignment yield representations closer to those of performant models (e.g., DinoV2)?
3. Does representational alignment provide evidence for the Platonic Representation Hypothesis?

Hypotheses:

1. A shared latent space exists where unimodal representations align via linear transformations.
2. Representational alignment improves similarity to performant models, as measured by kernel alignment metrics.
3. Multimodal alignment reflects representation convergence mechanisms suggested by the Platonic Representation Hypothesis.

Proposed Methodology

Dataset and Encoders

Dataset: Flickr30k (image-caption pairs) serves as a prototypical multimodal dataset. Each image is paired with its longest caption.

Encoders: Pre-trained unimodal models include:

- ResNet-18 for images, providing $\mathbb{R}^{d_x}$ features.
- GPT-2 for text, providing $\mathbb{R}^{d_y}$ features.

Mathematical Framework

The world generates multimodal data:

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

where $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are modalities (e.g., images, text, graphs). Pre-trained encoders extract unimodal features:

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

Linear Adapters: Modality-specific adapters map representations to a shared latent space $\mathbb{R}^{d_e}$:

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

These adapters are optimized using contrastive loss.

Contrastive Loss

The contrastive loss aligns representations from the same data point across modalities while separating unrelated representations:

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

Here, τ is a temperature hyperparameter.

Kernel Alignment

Kernel Definitions

1. **Unimodal Kernels:** Kernels induced by unimodal encoders:

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

2. **Aligned Multimodal Kernel:** Kernel for aligned representations:

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

3. **Performant Model Kernel:** Reference kernel from DinoV2:

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

Kernel Alignment Metric

A kernel alignment metric $m : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ measures the similarity between kernels. Comparisons:

1. Before training: $m(K_X, K_{\text{DinoV2}}), m(K_Y, K_{\text{DinoV2}})$.
2. After training: $m(K_{\text{repr}}, K_{\text{DinoV2}})$.

Key Hypothesis:

$$m(K_{\text{repr}}, K_{\text{DinoV2}}) > \text{avg}(m(K_X, K_{\text{DinoV2}}), m(K_Y, K_{\text{DinoV2}})).$$

Expected Contributions

1. A scalable framework for aligning pre-trained unimodal models into a shared latent space. 2. Empirical evidence for the Platonic Representation Hypothesis. 3. Insights into multimodal representation learning with frozen encoders.

References

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.