Draft Plan

Created @November 25, 2024 4:52 PM

Writeup for Initial Results, Plan, and Blog Post Outline

Introduction

Our hypothesis is that by aligning the embeddings of single-modality encoders ("unimodal representations") from different modalities using a simple transformation (e.g., linear adapters), we can produce a joint "multimodal representation" that better approximates the embeddings produced by a performant model (e.g., DinoV2). This performant model is assumed to yield representations closer to the underlying **platonic representation** of reality.

This project seeks to validate whether aligning representations from smaller, unimodal models using multimodal datasets can approximate more computationally expensive, large-scale models while achieving similar representation quality.

Our hypothesis rests on several assumptions:

1. Reality is an underlying (latent), potentially low-dimensional, generative model, and the optimal representation of this reality is the platonic representation.
    a. “joint distribution over events in the world that generate the data we observe”
2. More performant models (e.g., DinoV2) are closer to this platonic representation.

Related work

This work builds on:

1. **The Platonic Representation Hypothesis (Huh et al., 2024)**
    - Argues performant models converge toward a shared statistical model of reality in their representation spaces.
    - Identifies four drivers of this representational convergence:
        - Task-aligned objectives.
        - Multitask training.
        - Advanced architectures.
        - Multimodal data.
- **Existing multimodal alignment techniques:** Many state-of-the-art models (e.g., LLaVA) demonstrate that simple transformations like MLP projections can effectively align unimodal features into a joint multimodal space.

Relevant Studies

- **Huh et al., 2024**: Provided supporting evidence that representations are converging across modalities, meaning that models trained on *different* data modalities converge to one other.
- **Merullo et al. (2022):** Demonstrated that linear stitching (single linear projection) suffices to align vision models to LLMs, enabling cross-modal tasks like visual question answering.
- **Koh et al. (2023):** Extended this work to the opposite direction, aligning text representations to visual outputs.
- **Liu et al. (2023):** Developed LLaVA, demonstrating state-of-the-art results projecting visual features into a language model with a 2-layer MLP.
    - stitch pre-trained language and vision models together, effectively aligning visual features into language models
    - such multimodal alignment using simple projections (e.g MLPs) achieves strong performance.

Initial Results

Our initial experiments focused on two modalities, **images** and **text** (from the Flickr30k dataset), yielding the following insights:

1. **Alignment Works (with Caveats):**
    - Contrastively aligning image and text encoders produced a joint representation closer to the DinoV2 embeddings, partially validating our hypothesis.
    - Caveat: While the aligned representation exceeded the **average** similarity of unimodal kernels to DinoV2, it fell short of the **maximum** similarity (i.e., the image modality kernel).
2. **Cross-Modality Benefit:**
    - DinoV2 was trained only on images, yet the joint representation—involving text—became more similar to its embeddings. This suggests multimodal alignment enables representational convergence distinct from multitask training.
3. **Data Efficiency:**
    - Smaller datasets not only converged faster but achieved higher kernel alignment scores, highlighting the efficiency of contrastive learning.
4. **Kernel Alignment Metrics:**
    - m(text-modality kernel, DinoV2 kernel): Low similarity.
    - m(image-modality kernel, DinoV2 kernel): High similarity.
    - m(aligned multimodal kernel, DinoV2 kernel): Intermediate—higher than the average of unimodal kernels but skewed by the image modality.

Extensions (In-Scope for Final Project)

1. Downstream Task Evaluation

We will evaluate (e.g. on CIFAR10 classification) whether our joint representation is transferable to downstream tasks, focusing on:

- Replacing DinoV2's feature extractor with our aligned representation in-front of pre-trained heads (e.g., for ImageNet classification).
- Key hypotheses:
  - Performance will degrade slightly but remain better than chance and superior to an untrained DinoV2 model.
  - The aligned representation will outperform unimodal representations used in isolation.

Pretrained heads - Image classification

backbone	with registers	download
		ImageNet
ViT-S/14 distilled	❌	linear head (1 layer, 4 layers)
ViT-S/14 distilled	✅	linear head (1 layer, 4 layers)
ViT-B/14 distilled	❌	linear head (1 layer, 4 layers)
ViT-B/14 distilled	✅	linear head (1 layer, 4 layers)
ViT-L/14 distilled	❌	linear head (1 layer, 4 layers)
ViT-L/14 distilled	✅	linear head (1 layer, 4 layers)
ViT-g/14	❌	linear head (1 layer, 4 layers)
ViT-g/14	✅	linear head (1 layer, 4 layers)

Analysis Goals

Quantify Representation Quality

- Report final alignment scores for each kernel comparison to validate our hypothesis.

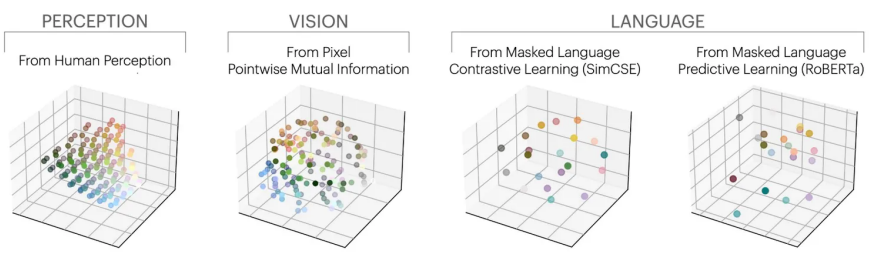
Mechanistic Insight

- Highlight how multimodal alignment enables cross-modality benefits, particularly for modalities like text that the performant model (DinoV2) wasn't explicitly trained on. Unfortunately our results do not show this!
- Investigate whether task-specific heads (e.g., VQA) in DinoV2 transform its image features into representations more akin to our joint multimodal alignment.
  - The idea is to compare our multimodal representation to that of the early layers of different pre-trained *heads* made DinoV2? The feature extractor part alone gives a representation space that is more similar to that of the unimodal image encoder than of the joint/multimodal alignment representation. But perhaps the task constraints introduced by the head of a task involving a

different modality (e.g. VQA) may transform the extracted features into representations that are more like that learn by our multimodal alignment training.

Proposed Visualizations:

1. **Kernel Alignment Over Epochs:**
   - Line plots tracking m(text-modality kernel), m(image-modality kernel), and m(aligned multimodal kernel) with DinoV2 kernels across training epochs.
     - This will highlight the improvement in alignment for each modality and the joint representation over the course of training.
   - Multidimensional scaling (MDS) plots for kernel matrices at key training milestones.



Kernels visualized with multidimensional scaling (i.e. a visualization where nearby points are similar according to the kernel, and far apart points are dissimilar). The language experiment here is a replication of Abdou et al. 2021.

2. **Visualizing Kernels:**
   - Select a small subset of images (e.g., 3-5) and visualize kernel for:
     - Text-modality kernel.
     - Image-modality kernel.
     - Aligned multimodal kernel.
     - DinoV2 kernel.
   - Visualize representational (dis)similarity for text, image, multimodal, and DinoV2 kernels for selected samples (3-5 images).

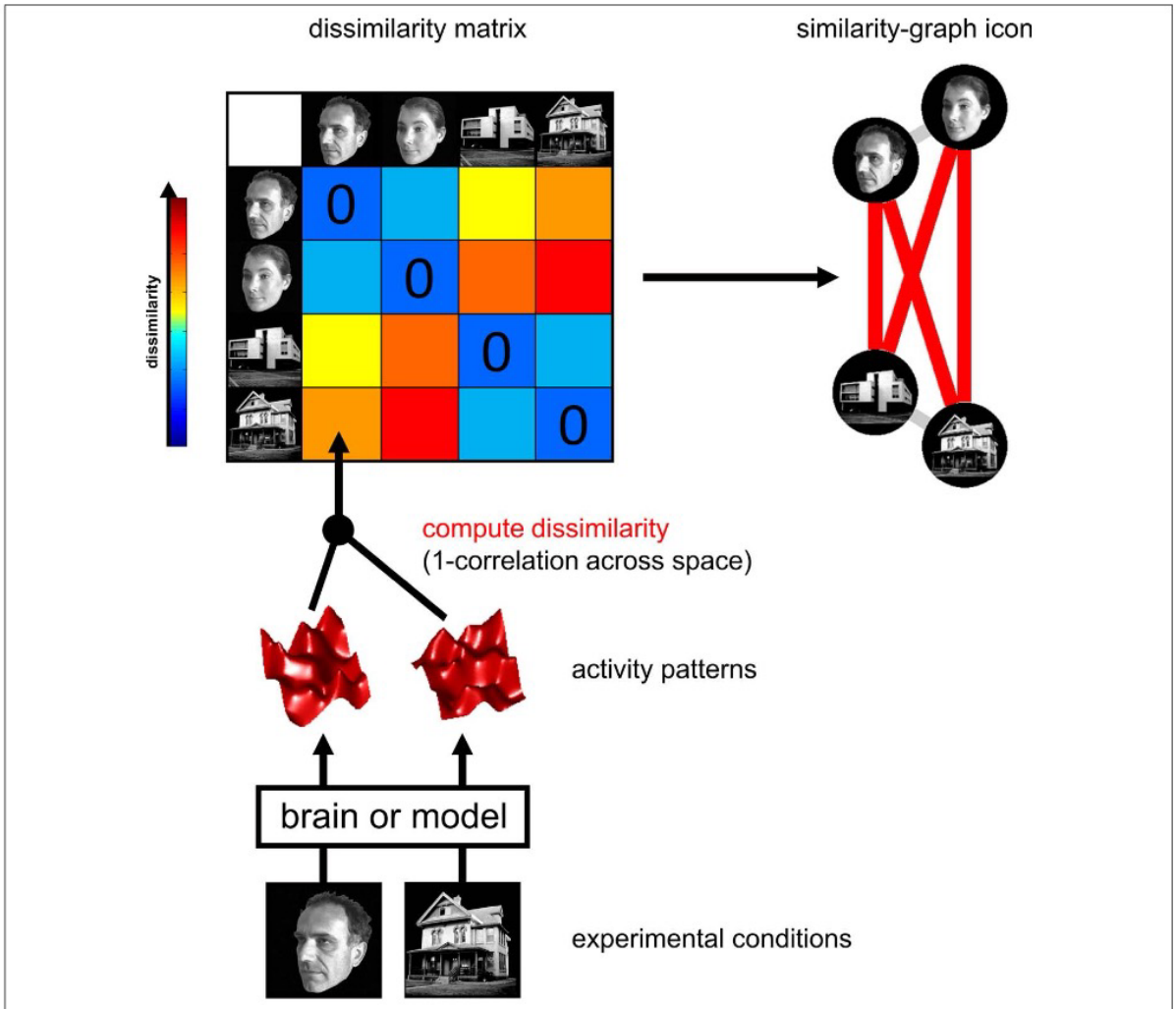
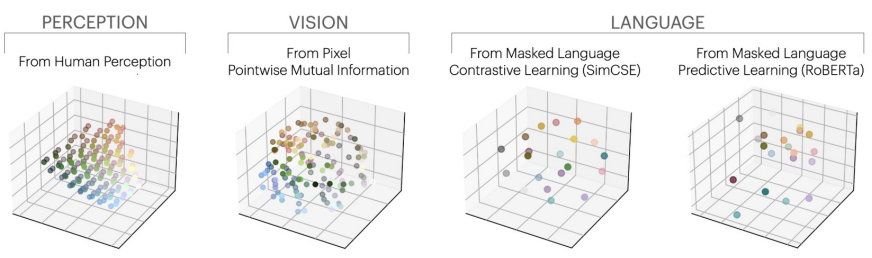


FIGURE 2 | Computation of the representational dissimilarity matrix. For each pair of experimental conditions, the associated activity patterns (in a brain region or model) are compared by spatial correlation. The dissimilarity between them is measured as 1 minus the correlation (0 for perfect correlation, 1 for no correlation, 2 for perfect anticorrelation). These dissimilarities for all pairs of conditions are assembled in the RDM. Each cell of the RDM, thus, compares the response patterns elicited by two images. As a consequence, an RDM is symmetric about a diagonal of zeros. To visualize the representation for a small number of conditions, we suggest the similarity-graph icon (top right, cf. **Figure 1**).

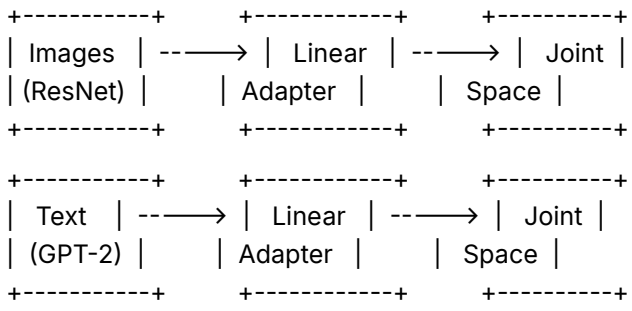
- or plot a multidimensional scaling visualization like they did in the **The Platonic Representation Hypothesis** paper:

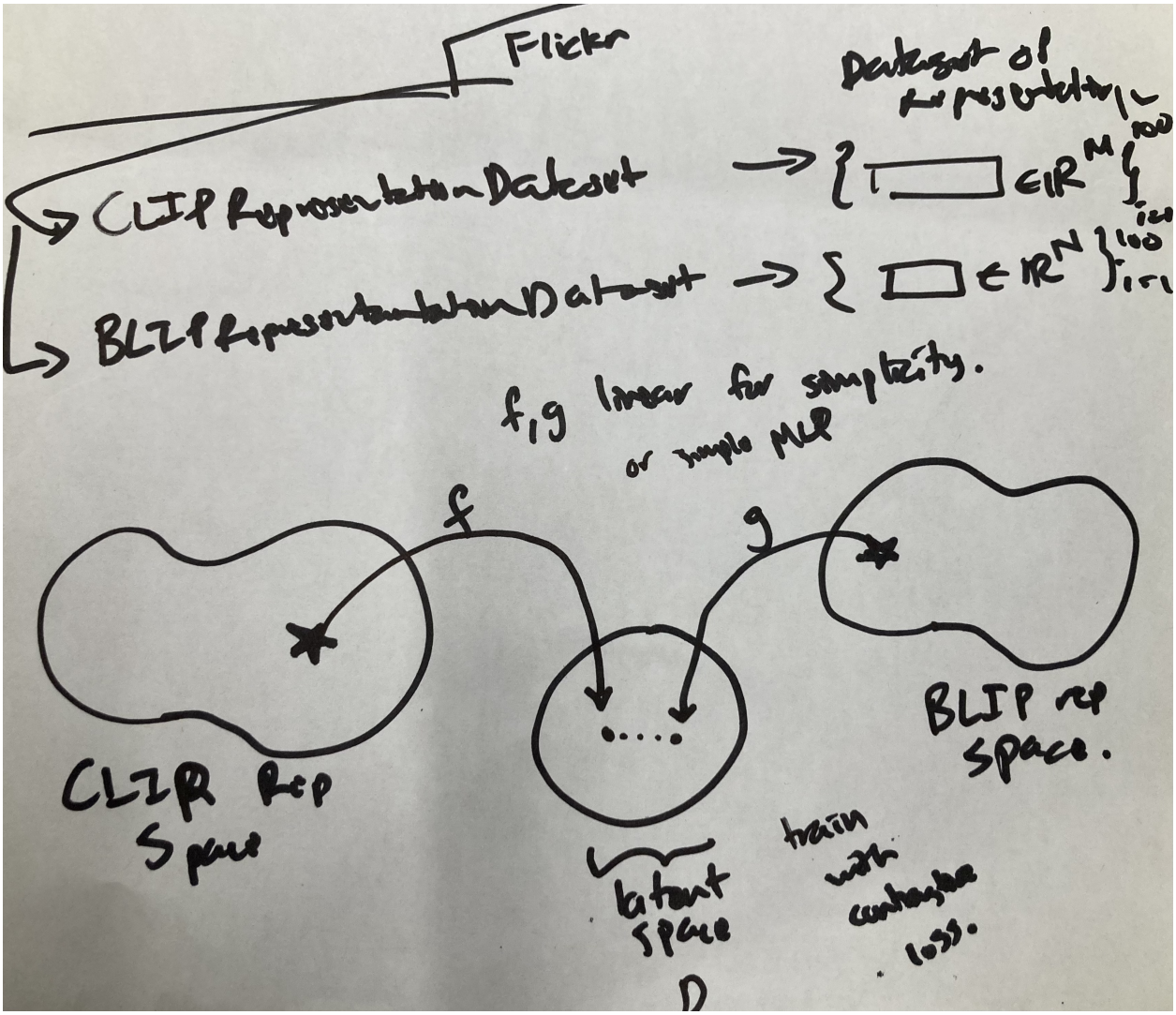


Kernels visualized with multidimensional scaling (i.e. a visualization where nearby points are similar according to the kernel, and far apart points are dissimilar). The language experiment here is a replication of Abdou et al. 2021.

- These visualizations should visually demonstrate whether aligned multimodal kernels are converging toward DinoV2 kernels.

3. **Representation PCA:**
    - Take the adapted unimodal encoder and get the representations for images, captions on curated test set of a few semantically categorized examples e.g. use prototype images from CIFAR100 with text “a photo of [caption]”.
    - Plot PCA colored by semantic level (category) labels.
    - Visualize PCA embeddings of representations at different training stages, highlighting the progression of alignment.
4. **Adapter Comparison:**
    - Bar charts comparing the kernel alignment scores and final training losses across adapter architectures (linear, MLP, bottleneck).
5. **Loss and Convergence Plots:**
    - Training loss and alignment metrics plotted against the number of epochs for different configurations (e.g., linear adapters vs. MLP adapters).
6. Add a **conceptual diagram** illustrating the process of aligning unimodal representations into a multimodal space.
    - Something like this sketch:





Concrete Plan for Course Project

Scope for Course Submission:

1. **Two Modality Prototype:**
   - Focus on aligning **images** (ResNet-18) and **text** (GPT-2) representations.
   - Use DinoV2 as the performant model to compute kernel alignment metrics.
   - Use Flickr30k as our “multimodal” dataset as it contains exactly those two modalities: images and text (captions)
2. **Experiments:**
   - Track alignment metrics for image, text, and aligned/joint representations against DinoV2.
   - Evaluate the effect of different adapter architectures (linear vs. MLP).
3. **Visualizations:**

- Include kernel alignment plots, heatmaps, and adapter comparison charts.

4. **Writing:**
   - Prepare a concise report detailing our hypothesis, methodology, initial results, and findings.

Future Directions (out-of-scope for final project deadline)

1. **Larger Datasets and Advanced Models:**
   - Scale experiments to include multimodal datasets with richer semantics (e.g., audio-video-caption datasets).
   - Benchmark against more advanced models like CLIP.
   - Investigate how alignment scales with the dimensionality of representation spaces.
2. **Mechanistic Insights:**
   - Investigate how multimodal alignment / contrastive learning affects inter-modality relationships.
   - Explore whether alignment leads to better downstream task performance (e.g., zero-shot classification, retrieval).
3. **Cross-Architecture Generalization:**
   - Test whether the aligned representation generalizes across architectures, e.g., transfer learned alignment from ResNet-GPT2 to another unimodal encoder pair.
4. **Adding a Third Modality**
   - If our prototype is promising, do the same thing above but adding the **audio** modality where the simple unimodal encoder could be a model like Wav2Vec2.
   - Now need a dataset that has those three modalities. VGG-Sound is a good candidate.
     - However, it doesn't seem straightforward to download this dataset.
     - Each line in the csv file has columns defined by:

       # YouTube ID, start seconds, label, train/test split.
   - Expanding beyond images and text, we will integrate audio as a third modality using the VGGSound dataset and Wav2Vec as the pre-trained audio encoder. The experiment will:
   - Test whether incorporating a third modality improves kernel alignment metrics.
   - Evaluate whether the joint multimodal representation achieves "the whole is greater than the sum of its parts," i.e.

     $$m(\text{aligned multimodal kernel, DinoV2-kernel}) > \max\{m(\text{image-modality kernel, DinoV2-kernel}), m(\text{text-modality kernel, DinoV2-kernel})\}$$

Blog Post Outline

Title:

"Aligning Modalities: Unlocking Multimodal Representations Without Large-Scale Training"

Sections:

1. **Introduction:**
    - Importance of representation learning.
    - Challenges of scaling large models.
    - Efficiency benefits of multimodal alignment.
2. **Hypothesis and Motivation:**
    - Explain the concept of platonic representations and why they matter.
    - Discuss efficiency benefits of leveraging unimodal models and multimodal datasets.
3. **Methodology:**
    - Our method of aligning smaller encoders to approximate performant models.
    - Describe dataset setup, contrastive alignment, and kernel alignment metrics.
4. **Results:**
    - Alignment success and caveats.
    - Include all proposed visualizations (alignment plots, representation PCA, heatmaps).
5. **Implications:**
    - Discuss broader significance for multimodal AI.
6. **Future Directions:**
    - Outline potential extensions (scaling dataset sizes, additional modalities, downstream tasks).
7. **Takeaways:**
    - Highlight key findings: successful alignment of multimodal representations, potential for replacing large-scale models in some scenarios.
    - Emphasize the simplicity and efficiency of the approach.
    - Encourage exploration of multimodal alignment of many more modalities.