Exploring the Platonic Representation Hypothesis Beyond In-Distribution Data

Aryasomayajula Ram Bharadwaj
Independent Researcher
ram.bharadwaj.arya@gmail.com

20th Oct 2024

Abstract

The Platonic Representation Hypothesis (PRH) [1] posits that neural networks, despite being trained on different objectives and datasets, converge toward a shared statistical model of reality. This paper extends the investigation of PRH to out-of-distribution (OOD) settings, using ImageNet-O as a benchmark, and contrasts the findings with results from in-distribution data and random noise. Our analysis reveals that while PRH holds in OOD scenarios, it breaks down with purely random data, suggesting that a common underlying structure in the data is essential for representational alignment.

1 Introduction

The PRH suggests that models trained across various modalities and objectives converge toward a shared representation of reality [1]. Inspired by philosophical concepts like Plato's Allegory of the Cave, PRH proposes that neural networks are progressively learning an idealized representation of the world. This convergence has been observed across domains such as vision and language models, with evidence showing that larger models better align their internal representations [1]. However, it remains unclear whether such convergence persists when models encounter data significantly different from their training distribution.

To address this, we extend PRH's analysis to out-of-distribution (OOD) settings using the ImageNet-O dataset, which contains samples outside the distribution of ImageNet [2]. We also include random noise as a contrasting dataset to examine the boundaries of representational alignment.

2 Methodology

We assess representational alignment using various metrics, such as mutual k-NN and CKNN-A, across three types of data:

1. **In-distribution data**: Data that aligns with the pretraining dataset (e.g., Places365's validation set).
2. **Out-of-distribution data**: ImageNet-O, designed with images that are outliers to the ImageNet distribution [2].
3. **Random noise**: Purely random images to test the limits of PRH.

The alignment between different models is measured using Spearman's rank correlation, focusing on how similar the models' distance measures are for given data points. Throughout the experiments, 17 different Vision transformers spanning different sizes and trained data are used. Mutual k-Nearest Neighbor Alignment is measured in all the three settings.

3 Results

3.1 Alignment on In-Distribution Data

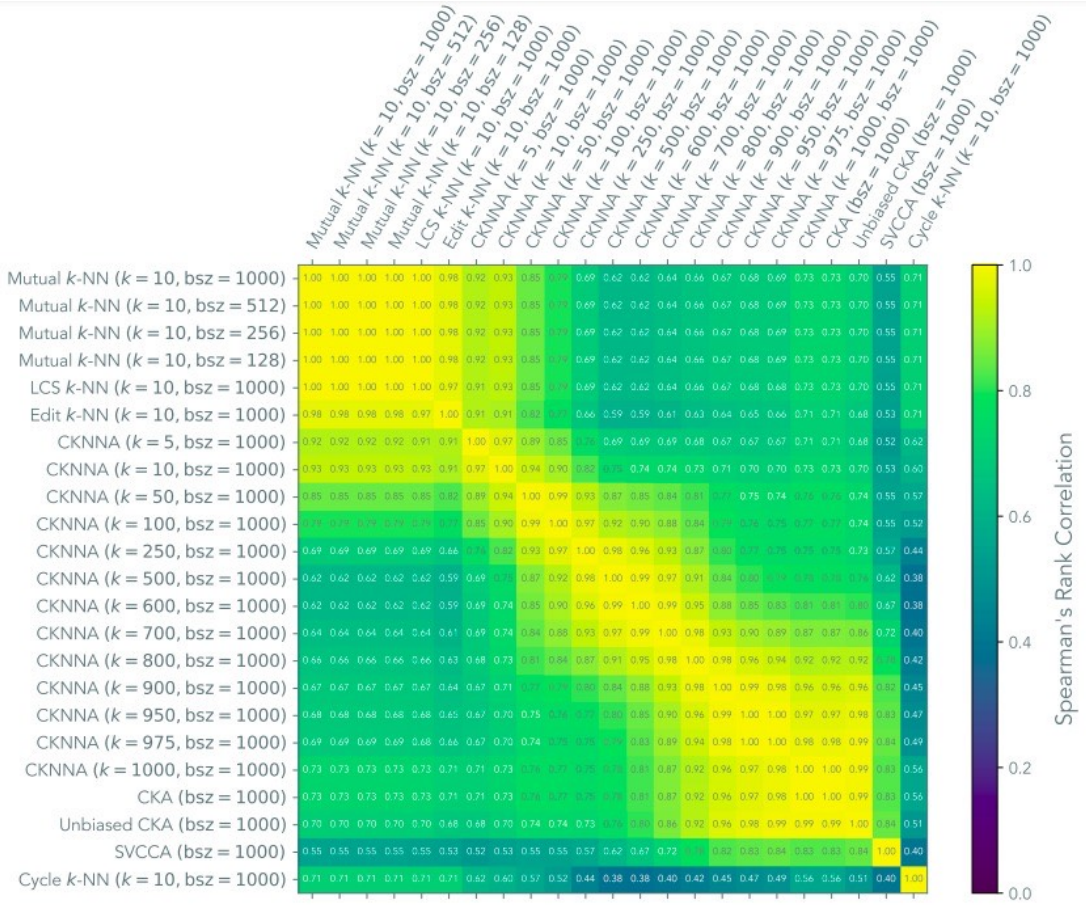

Figure 12. **Vision-vision alignment measured with various metrics.** Spearman's rank correlation among different metrics and batch sizes (bsz) when used to measure alignment among 78 vision models (see Appendix C.1 for details of these models). All p-values are below 2.24×10^{-105}. Our vision-vision analysis in Figure 2 is based on the first metric (Mutual k-NN with $k = 10$ and bsz $= 1000$).

Figure 1: Vision-vision alignment measured on Places365's validation dataset using various metrics. High Spearman's rank correlation among different metrics indicates alignment among vision models.

3.2 Alignment on ImageNet-O

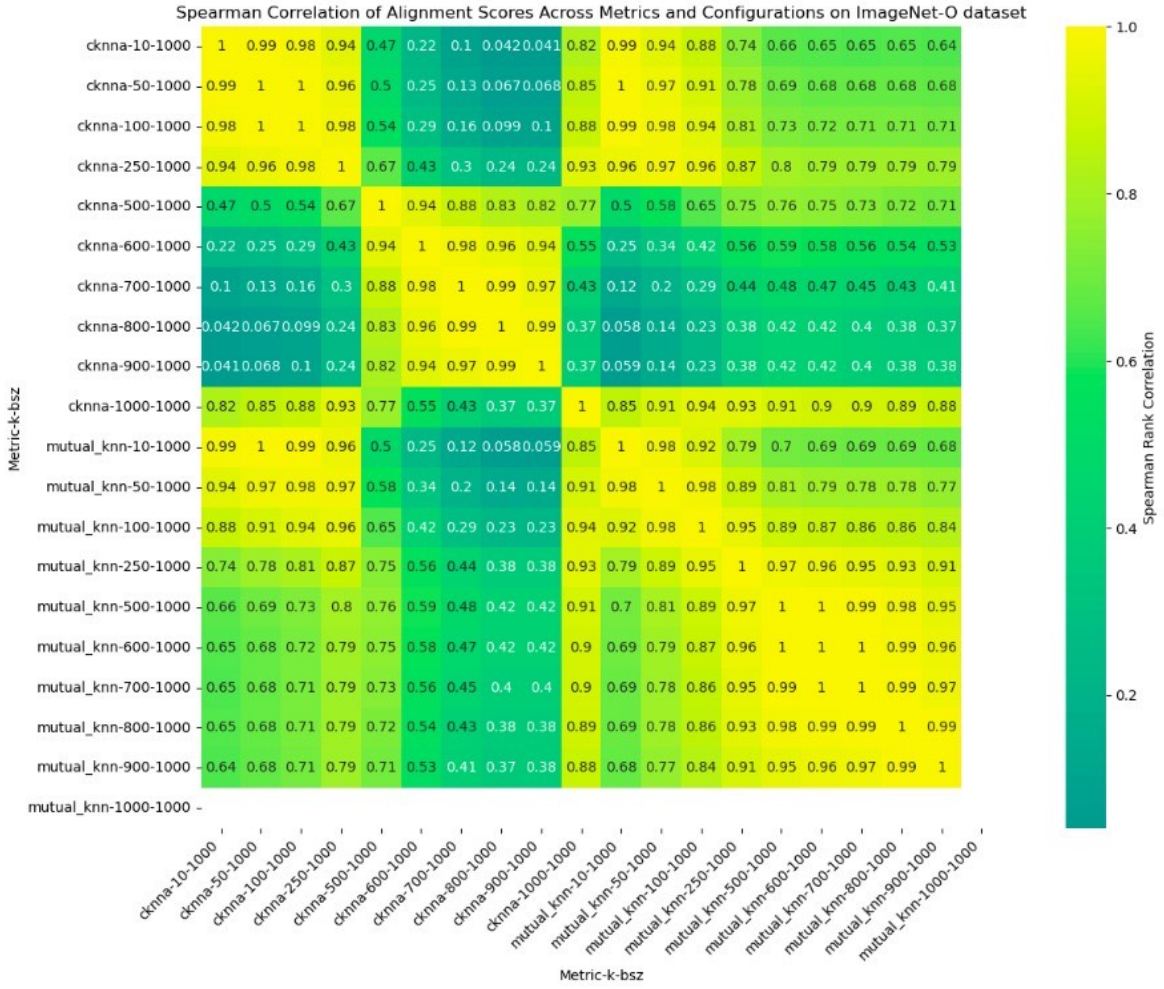

Figure 2: Alignment of vision-models on ImageNet-O dataset. The Spearman correlation suggests that even with outlier data, models maintain a shared statistical representation, albeit with higher prediction errors.

3.3 Alignment on Random Data

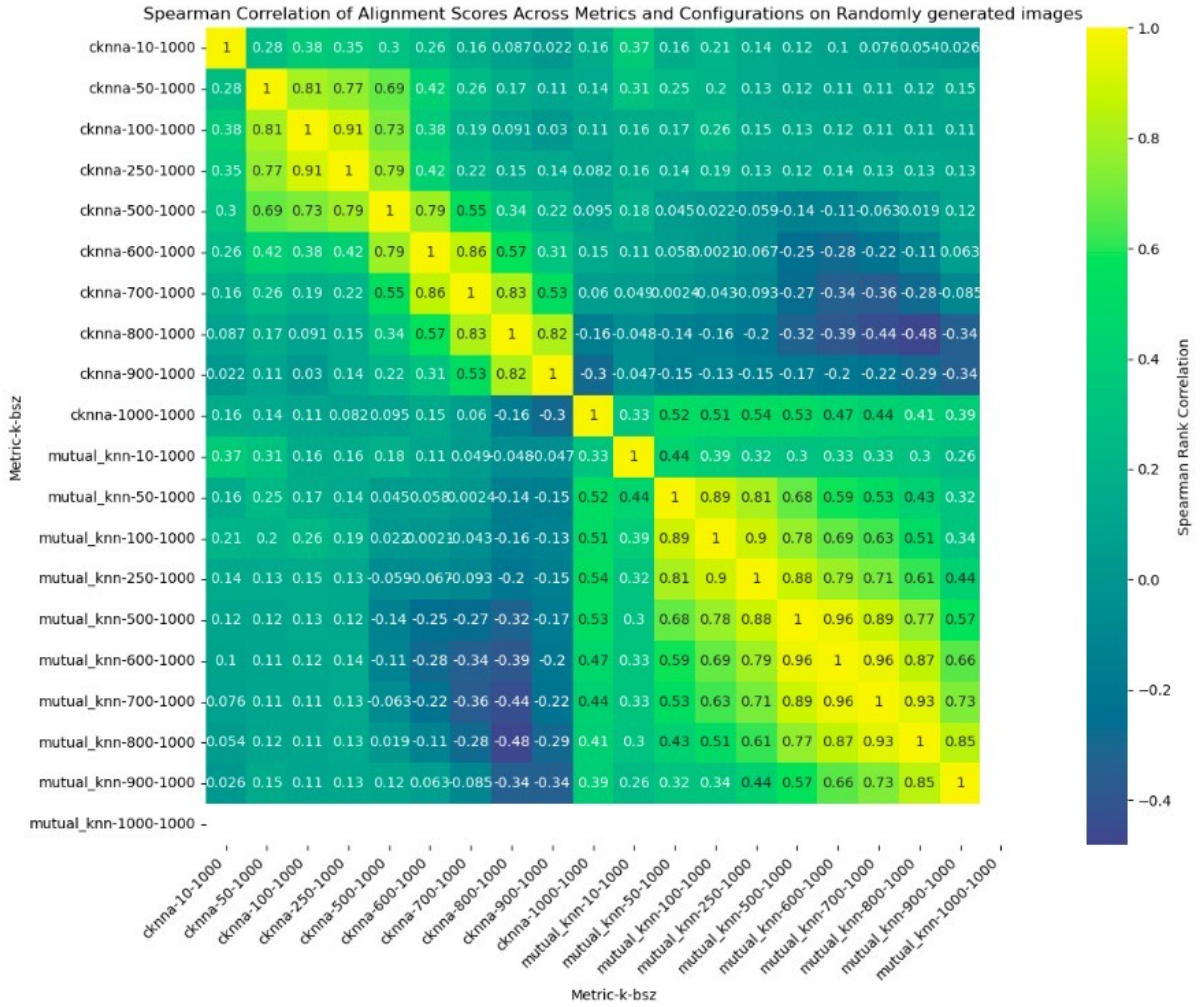

Figure 3: Spearman correlation of alignment scores on random noise data. The lower correlation values indicate a breakdown in representational alignment, suggesting that random data lacks the underlying structure needed for PRH.

4 Discussion

Our findings show that PRH extends beyond in-distribution data to OOD settings, where models exhibit alignment in their internal representations. On ImageNet-O, models align even in their predictions even when they produce incorrect results with high confidence, revealing that models fail in a predictable manner on OOD data. This suggests that they rely on a shared interpretation of the input space, even when the data diverges from what they have seen during training.

However, this alignment does not extend to purely random data, where the lack of structure prevents models from developing a shared statistical understanding. This distinction emphasizes that the presence of meaningful underlying structure is a prerequisite for representational convergence, aligning with PRH's notion of a shared representation of reality.

5 Conclusion

The results validate the Platonic Representation Hypothesis in OOD scenarios but highlight its limitations with random noise. Our analysis underscores the need for structured data in achieving representational alignment and suggests new directions for exploring the boundaries of PRH. Future work could focus on extending PRH to other modalities and finding conditions where the hypothesis might be challenged or refuted.

The code used for the experiments and analysis in this paper is available on GitHub here.

References

[1] Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic Representation Hypothesis. *International Conference on Machine Learning*.

[2] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural Adversarial Examples. *CVPR*. arXiv:2405.07987.