The Platonic Representation Hypothesis

Minyoung Huh [*1] **Brian Cheung** [*1] **Tongzhou Wang** [*1] **Phillip Isola** [*1]

MIT, Arxiv'24

Key Conclusion

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

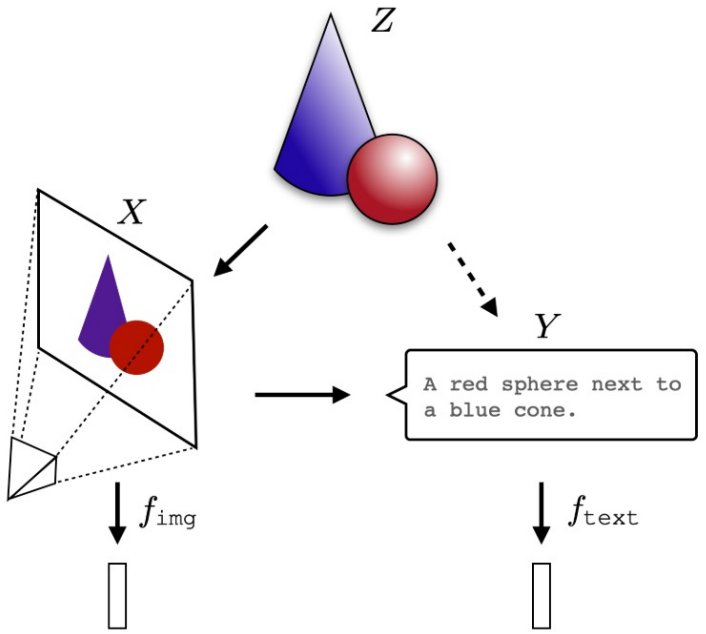


Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.

Why?
How?
What?

Preliminaries

- A **representation** is a function $f : \mathcal{X} \to \mathbb{R}^n$ that assigns a feature vector to each input in some data domain $\mathcal{X}$.
- A **kernel**, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, characterizes how a representation measures distance/similarity between datapoints. $K(x_i, x_j) = \langle f(x_i), f(x_j) \rangle$, where $\langle \cdot , \cdot \rangle$ denotes inner product, $x_i, x_j \in \mathcal{X}$ and $K \in \mathcal{K}$.
- A **kernel-alignment metric**, $m : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$, measures the similarity between two kernels, *i.e.*, how similar is the distance measure induced by one representation to the distance measure induced by another. Examples include Centered Kernel Distance (CKA) (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), and nearest-neighbor metrics (Klabunde et al., 2023).

In our experiments, we use a *mutual nearest-neighbor metric* that measures the mean intersection of the k-nearest neighbor sets induced by two kernels, K_1 and K_2, normalized by k. This metric is a variant of those proposed

Preliminaries

A. Mutual k-Nearest Neighbor Alignment Metric

For two models with representations f, g the mutual k-nearest neighbor metric measures the average overlap of their respective nearest neighbor sets. In this section, we refer to this metric as m_{NN}, which we will formally define below.

For cross-modal domains, define $(x_i, y_i) \in \mathcal{X}$ as a sample from the data distribution $\mathcal{X}$ (*e.g.* image-caption dataset). For the single domain alignment measurements, the samples are equivalent $x_i = y_i$ (*e.g.*, images for vision, and text for language). Let $\{x_i, y_i\}_{i=1}^{b}$ be the corresponding mini-batch sampled from this data distribution. Then given two model representations f and g the corresponding features are: $\phi_i = f(x_i)$ and $\psi_i = f(y_i)$, where the collection of these features are denoted as $\Phi = \{\phi_1, \ldots, \phi_b\}$ and $\Psi = \{\psi_1, \ldots, \psi_b\}$. Then for each feature pair (ϕ_i, ψ_i), we compute the respective nearest neighbor sets $\mathcal{S}(\phi_i)$ and $\mathcal{S}(\psi_j)$.

$$d_{\mathsf{knn}}(\phi_i, \Phi \setminus \phi_i) = \mathcal{S}(\phi_i) \tag{9}$$

$$d_{\mathsf{knn}}(\psi_i, \Psi \setminus \psi_i) = \mathcal{S}(\psi_j) \tag{10}$$

where d_{knn} returns the set of indices of its k-nearest neighbors. Then we measure its average intersection via

$$m_{\mathsf{NN}}(\phi_i, \psi_i) = \frac{1}{k} |\mathcal{S}(\phi_i) \cap \mathcal{S}(\psi_j)| \tag{11}$$

where $|\cdot|$ is the size of the intersection.

Representations are converging

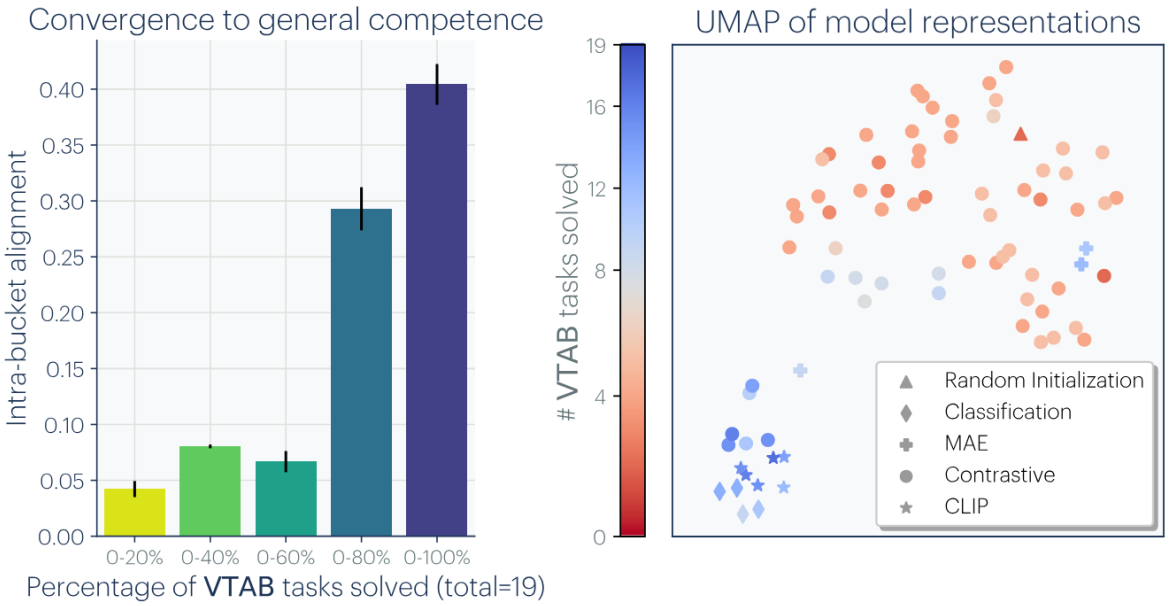


Figure 2. **VISION models converge as COMPETENCE increases:** We measure alignment among 78 models using mutual nearest-neighbors on Places-365 (Zhou et al., 2017), and evaluate their performance on downstream tasks from the Visual Task Adaptation Benchmark (VTAB; Zhai et al. (2019)). **LEFT:** Models that solve more VTAB tasks tend to be more aligned with each other. Error bars show standard error. **RIGHT:** We use UMAP to embed *models* into a 2D space, based on distance $\triangleq -\log(\text{alignment})$. More competent and general models (blue) have more similar representations.

Different models, with different architectures and objectives, can have aligned representations.

Representations are converging

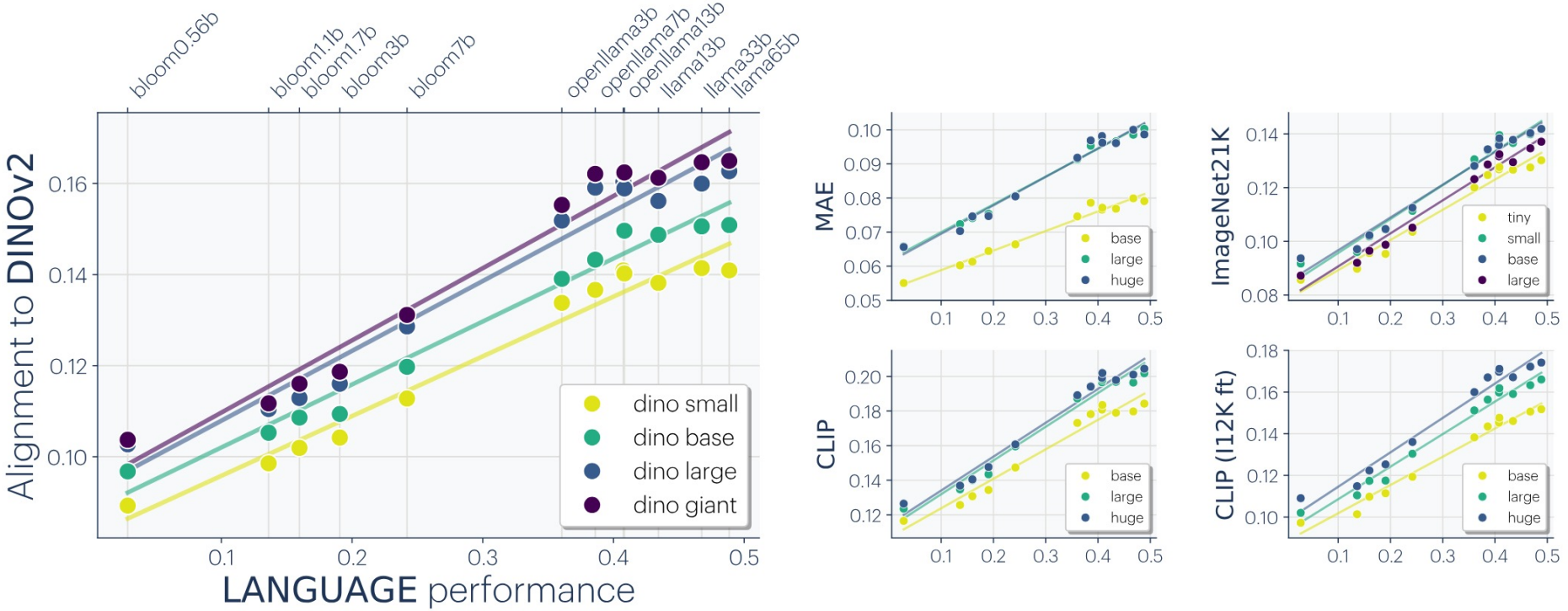


Figure 3. **LANGUAGE and VISION models align:** We measure alignment using mutual nearest-neighbor on the Wikipedia caption dataset (WIT) (Srinivasan et al., 2021). The x-axis is the language model performance measured over 4M tokens from the OpenWebText dataset (Gokaslan & Cohen, 2019) (see Appendix B for plots with model names). We measure performance using $1 - \texttt{bits-per-byte}$, where `bits-per-byte` normalizes the cross-entropy by the total bytes in the input text string. The results show a linear relationship between language-vision alignment and language modeling score, where a general trend is that more capable language models align better with more capable vision models. We find that CLIP models, which are trained with explicit language supervision, exhibit a higher level of alignment. However, this alignment decreases after being fine-tuned on ImageNet classification (labeled CLIP (I12K ft)).

Representations are converging across modalities.

Dinov2: Learning robust visual features without supervision, 2023.

Representations are converging

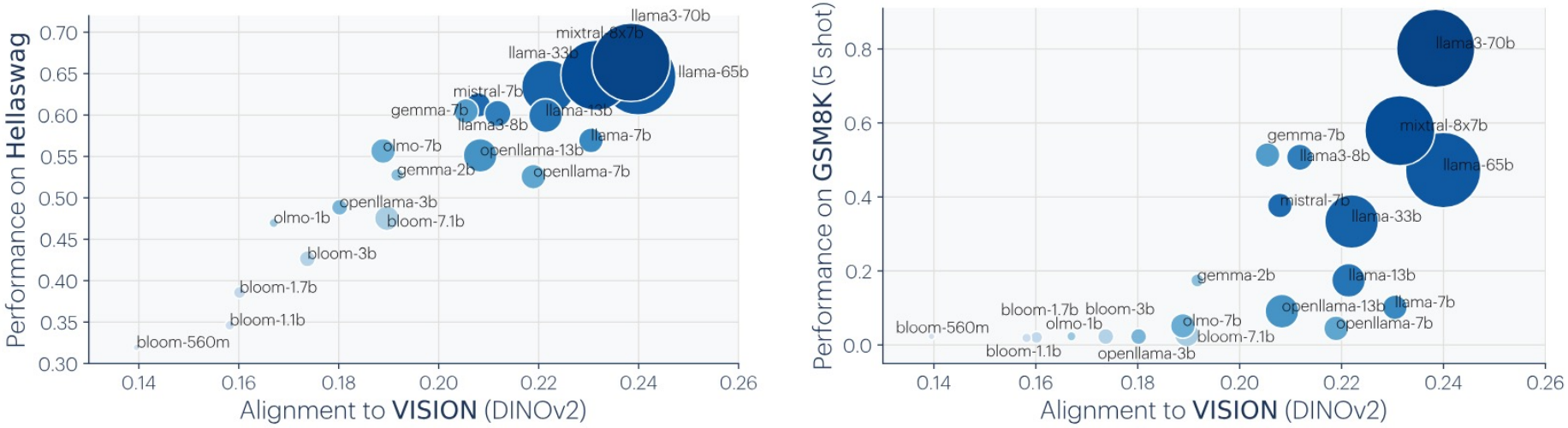


Figure 4. **Alignment predicts downstream performance:** We visualize correlation between LLM alignment score to DINOv2 (Oquab et al., 2023) and downstream task performance on Hellaswag (common-sense) (Zellers et al., 2019) and GSM8K (math) (Cobbe et al., 2021). LLMs are plotted with radii proportional to the size of the model, and color-coded by their rank order in language modeling scores ($1 - \texttt{bits-per-byte}$). We observe that models aligned more closely with vision also show better performance on downstream language tasks. For Hellaswag, there is a linear relationship with alignment score, while GSM8K exhibits an "emergence"-esque trend.

If models are converging towards a more accurate representation of reality, we expect that alignment should correspond to improved performance on downstream tasks.

Why are representations converging?

Modern machine learning models are generally trained to minimize the empirical risk with possible implicit and/or explicit regularization:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \text{dataset}} [\mathcal{L}(f, x)] + \mathcal{R}(f)$$

trained model: f^*; training objective: $\mathcal{L}$; function class: $\mathcal{F}$; regularization: $\mathcal{R}$

In the following sections, we lay out how each colored component in this optimization process potentially plays a role in facilitating representational convergence.

3.1. Convergence via Task Generality

3.2. Convergence via Model Capacity

3.3. Convergence via Simplicity Bias

Why are representations converging?

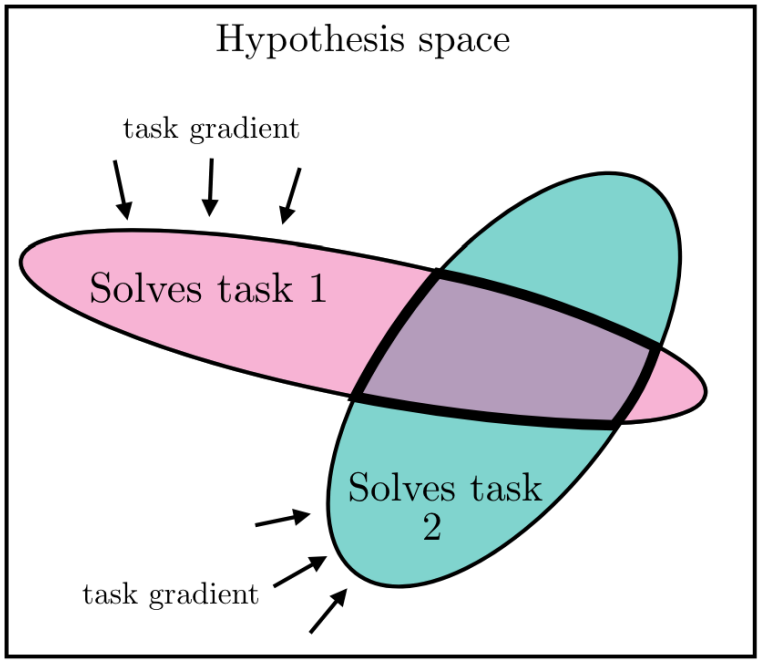


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

The Multitask Scaling Hypothesis

There are fewer representations that are competent for N tasks than there are for $M < N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions.

Why are representations converging?

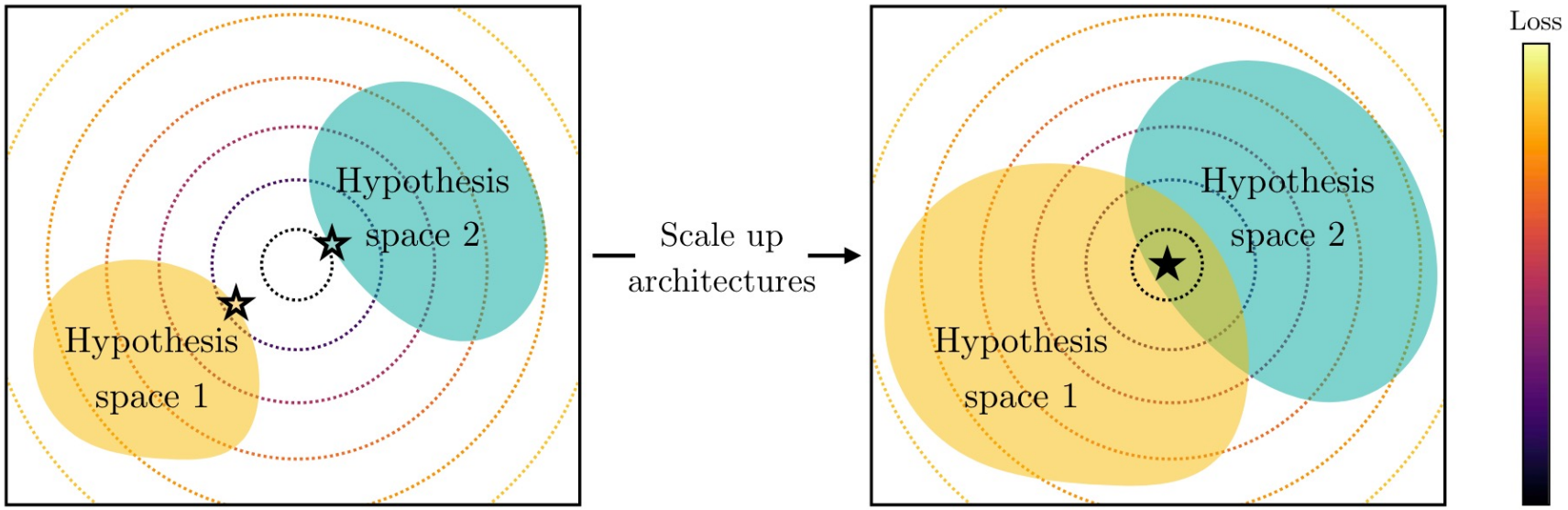
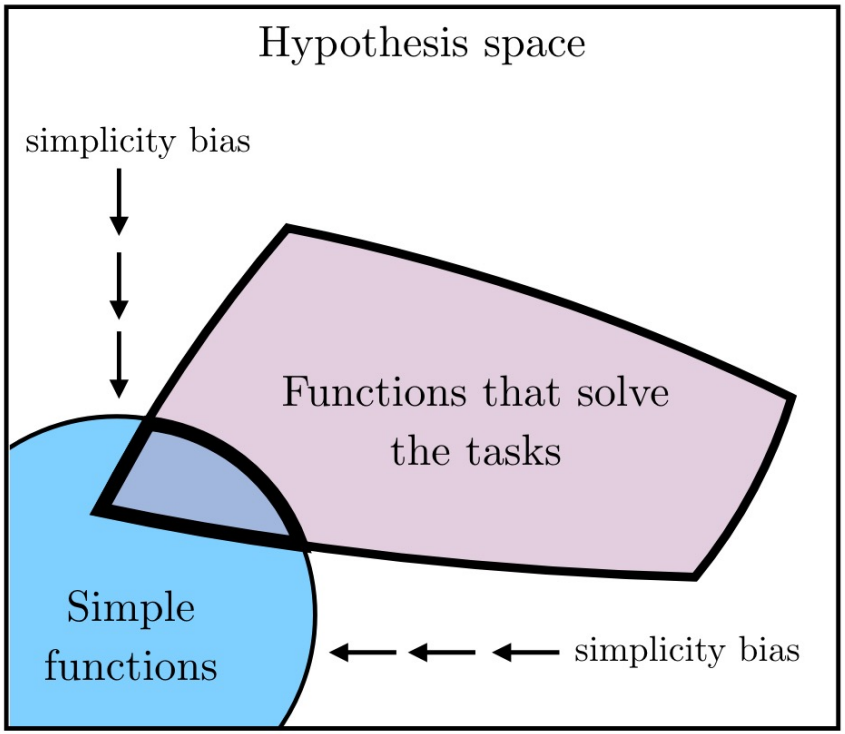


Figure 5. **The Capacity Hypothesis:** If an optimal representation exists in function space, larger hypothesis spaces are more likely to cover it. **LEFT:** Two small models might not cover the optimum and thus find *different* solutions (marked by outlined ☆). **RIGHT:** As the models become larger, they cover the optimum and converge to the same solution (marked by filled ★).

The Capacity Hypothesis

Bigger models are more likely to converge to a shared representation than smaller models.

Why are representations converging?



The Simplicity Bias Hypothesis

Deep networks are biased toward finding simple fits to the data, and the bigger the model, the stronger the bias. Therefore, as models get bigger, we should expect convergence to a smaller solution space.

Figure 7. **The Simplicity Bias Hypothesis:** Larger models have larger coverage of all possible ways to fit the same data. However, the implicit simplicity biases of deep networks encourage larger models to find the simplest of these solutions.

What representation are converging to?

- Task and data pressures, combined with increasing model capacity can lead to representation convergence.

- What is the endpoint of all this convergence.

Our central hypothesis, stated in Figure 1, is that the representation we are converging toward is a statistical model of the underlying reality that generates our observations. Consistent with the multitask scaling hypothesis, such a representation would naturally be useful toward many tasks (or at least toward any task grounded in reality). Additionally, this representation might be relatively simple, assuming that scientists are correct in suggesting that the fundamental laws of nature are indeed simple functions (Gell-Mann, 1995), in line with the simplicity bias hypothesis.

What representation are converging to?

4.1. An idealized world

We consider a world that works as follows, consistent with the cartoon in Figure 1. The world consists of a sequence of T discrete events, denoted as $\mathbf{Z} \triangleq [z_1, \ldots, z_T]$, sampled from some unknown distribution $\mathbb{P}(\mathbf{Z})$. Each event can be observed in various ways. An observation is a bijective, deterministic function $\mathsf{obs} : \mathcal{Z} \to \cdot$ that maps events to an arbitrary measurement space, such as pixels, sounds, mass, force, torque, words, etc. Later, in Section 6, we discuss limitations and potential extensions to continuous and unbounded worlds, and stochastic observations, that could yield a model that better reflects real learning scenarios.

This convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality (platonic representation).

What representation are converging to?

$$P_{\mathsf{coor}}(x_a, x_b) \quad \propto \sum_{(t,t')\,:\,|t-t'|\le T_{\mathsf{window}}} \mathbb{P}(X_t = x_a, X_{t'} = x_b).$$

$$\langle f_X(x_a), f_X(x_b)\rangle \approx \log \frac{\mathbb{P}(\mathsf{pos} \mid x_a, x_b)}{\mathbb{P}(\mathsf{neg} \mid x_a, x_b)} + \tilde{c}_X(x_a) \quad (3)$$

$$= \log \frac{P_{\mathsf{coor}}(x_a \mid x_b)}{P_{\mathsf{coor}}(x_a)} + c_X(x_a) \quad (4)$$

$$= K_{\mathsf{PMI}}(x_a, x_b) + c_X(x_a), \quad (5)$$

where K_{PMI} is the pointwise mutual information (PMI) kernel, and $c_X(x_a)$ is constant in x_b. We note that is a common

This analysis suggests that certain representation learning algorithms may boil down to a simple rule: *find an embedding in which similarity equals PMI*. We note that this idea is consistent with prior works that have used PMI as a similarity measure for clustering in vision and language

What are the implications of convergence?

Scaling is sufficient, but not necessarily efficient Our arguments are roughly in line with the claim that "scale is all you need" to reach high levels of intelligence. We have argued that as resources are scaled (# parameters, # datapoints, # flops), representations are converging, regardless of other modeling choices and even data modality. Does this mean that scale is all that matters? Not quite: different methods can scale with different levels of *efficiency* (Hestness et al., 2017; Kaplan et al., 2020), and successful methods must still satisfy some general requirements (*e.g.*, be a consistent estimator, model pairwise statistics of $\mathbb{P}(\mathbf{Z})$).

What are the implications of convergence?

Training data can be shared across modalities Suppose you have access to N images and M sentences, and want to learn the best representation. If there is indeed a modality-agnostic platonic representation, then the image data should help find it, and so should the language data. The implication is that if you want to train the best vision model, you should train not just on N images but also on M sentences. This is already becoming common practice (Achiam et al., 2023; Radford et al., 2021). Many vision models are fine-tuned from pre-trained LLMs. The other direction is less common, but also is implied by our hypothesis: if you want to build the best LLM, *you should also train it on image data*. Indeed, Achiam et al. (2023) claim evidence that this is true, where training on images improved performance on text. In theory, there should be some conversion ratio: a pixel is worth a words for training LLMs, and a word is worth b pixels for training vision models.

What are the implications of convergence?

Scaling may reduce hallucination and bias A prominent shortcoming of current LLMs is their propensity to hallucinate, or output false statements. If models are indeed converging toward an accurate model of reality, and scale powers this convergence, then we may expect hallucinations to decrease with scale. Of course, our hypothesis is conditioned on the training data for future models constituting a sufficiently lossless and diverse set of measurements. This may not come to pass, but it is an implication of our hypothesis worth pointing out. A similar argument can be made about certain kinds of bias. It has been shown that large models can exacerbate existing biases present in their training data (Hall et al., 2022). Our hypothesis implies that, while this may be true, we should expect *larger* models to amplify bias *less*. This does not mean bias will be removed, rather that the model's biases will more accurately reflect the data's biases, rather than exacerbating them.

What are the implications of convergence?

Different modalities may contain different information.

作者强调了在处理非双射观察和抽象概念时，信息量和模型容量的重要性。作者提出了一个更细化的假设版本，即只有在输入信号信息量足够高且模型容量足够大时，不同模型才能收敛到相同的表示。当条件不满足时，表示的对齐程度受到互信息和模型容量的限制。此外，通过初步测试，作者发现信息量越高的输入信号，其表示对齐效果越好。

下面给出一个简单的例子来说明

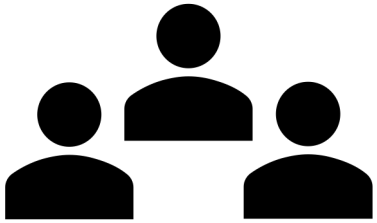
假设我们有两个小队，各自要完成两项任务：看一段视频并描述其中的内容，和读一段文字并画出其中的情景。

1. 高信息量和高容量的情况下：
   视频非常清晰，文字描述非常详细（高信息量的输入），小队的成员都很擅长这两项任务（高模型容量）。
   小队A和小队B描述视频内容时，文字描述会非常相似；读完文字并画出的情景图也会非常相似。
2. 低信息量或低容量的情况下：
   视频模糊不清，文字描述简略（低信息量的输入），小队的成员经验不足（低模型容量）。
   小队A和小队B描述视频内容时，文字描述会有很大差异；读完文字并画出的情景图也会差别很大。

Not all representations are presently converging.

尽管视觉和语言这两种模态的表示正在逐渐收敛，但其他模态尚未达到同样的水平。例如，在机器人领域，尚未形成一种像图像和文本那样标准化的表示世界状态的方法。这表明表示的收敛在不同领域中存在差异。收敛的挑战与限制主要体现在硬件限制与数据不足两方面。尽管存在这些挑战，作者们预期其他模态在未来可能会遵循类似的趋势，逐渐实现表示的收敛。

Discussion



Observations

Reality world

human

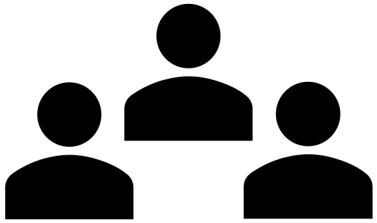
Conclude

- Math
- Physics
- Biology
- Chemistry
- ……

Conclude

- Taylor function, Calculus, …
- $F = ma, \quad E = mc^2$, …
- ……

Discussion



human

Observations

Reality world

Conclude

- Math
- Physics
- Biology
- Chemistry
- ……

Conclude

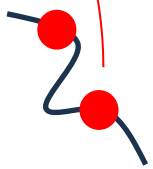
- Taylor function, Calculus, …
- $F = ma,\ \ E = mc^2$, …
- ……

Observe some easy pattens

Projection

Plato's ideal world

Complex pattens:



Discussion

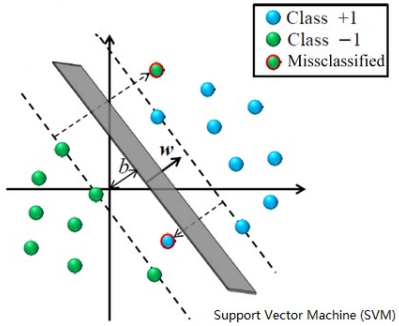
- Math
- Physics
- Biology
- Chemistry
- ……

Conclude →

Math Tools

- Taylor function, Calculus, …
- $F = ma,\ \ E = mc^2$, …
- ……

Current tools are limited.



Fitting

Plato's ideal world

Complex pattens:

Existing | Explanation

Discussion

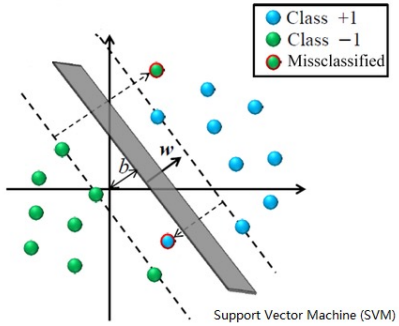
- Math
- Physics
- Biology
- Chemistry
- ……

Conclude →

Math Tools

- Taylor function, Calculus, …
- $F = ma,\ \ E = mc^2$, …
- ……

Current tools are limited.



Fitting

Plato's ideal world

Complex pattens:

Existing | Explanation

AI4S may be is right?

Thanks