柏拉图表征假说 - 概要分析

作者：Calvin
QQ: 179209347
Mail: 179209347@qq.com

介绍

笔记简介：

- 面向对象：深度学习初学者
- 依赖课程：**线性代数，统计概率**，优化理论，图论，离散数学，微积分，信息论

知乎专栏：

https://zhuanlan.zhihu.com/p/693738275

Github & Gitee 地址：

https://github.com/mymagicpower/AIAS/tree/main/deep_learning

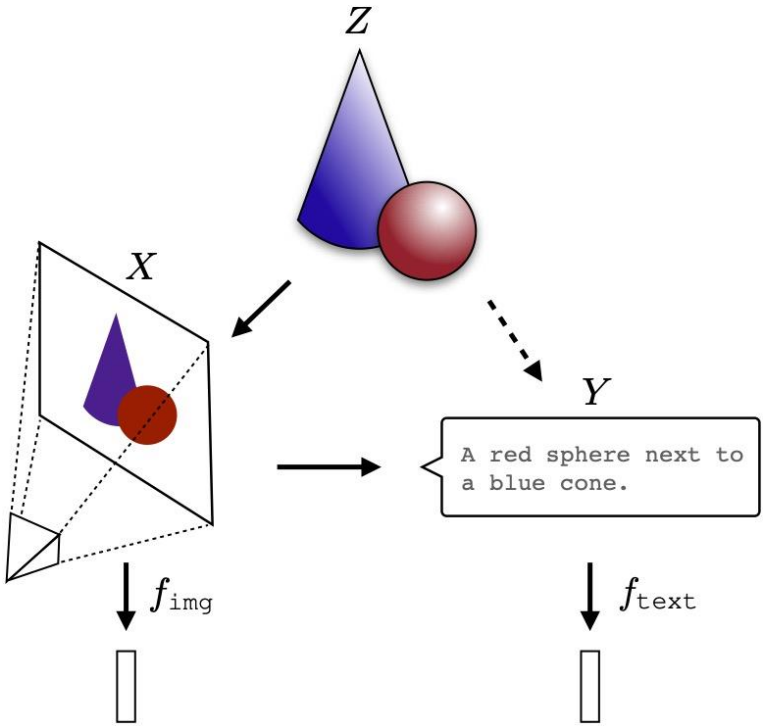
https://gitee.com/mymagicpower/AIAS/tree/main/deep_learning

*** 版权声明：**

- 仅限用于个人学习
- 禁止用于任何商业用途

柏拉图表征假说（Platonic Representation Hypothesis）

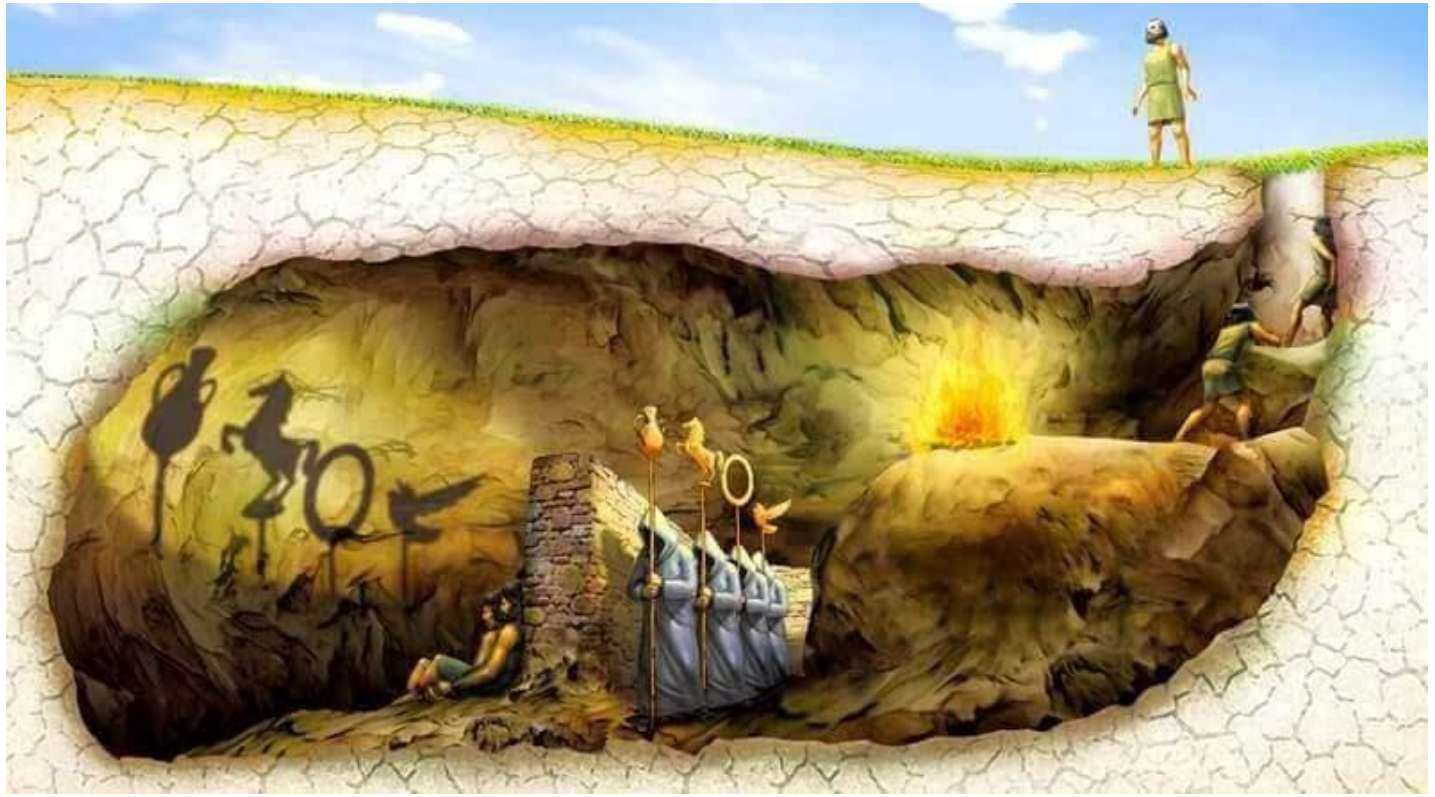
柏拉图表象假说（Platonic Representation Hypothesis）是一个理论概念：
它认为在人工智能（AI）模型中，尤其是深度神经网络，随着模型规模的扩大和训练任务的多样化，**不同的模型在表示数据的方式上越来越趋于一致**。这种趋同指向一个共享的统计模型，这个模型能够捕捉到现实世界的基本结构，类似于古希腊哲学家柏拉图关于理想现实的概念。



世界（Z）可以用许多不同的方式来看待，如：图像（X），文本（Y）等。在每种模态上学习的表征将收敛到 Z 的类似表征。
这一假说与柏拉图的洞穴寓言相联系，其中真实世界被视为理想的形式，而我们通过感官体验到的是这些理想形式的影子或映射。

柏拉图洞穴寓言（Plato's Allegory of the Cave）

柏拉图想象了一个“理想”的现实，我们的观察只是它的影子。

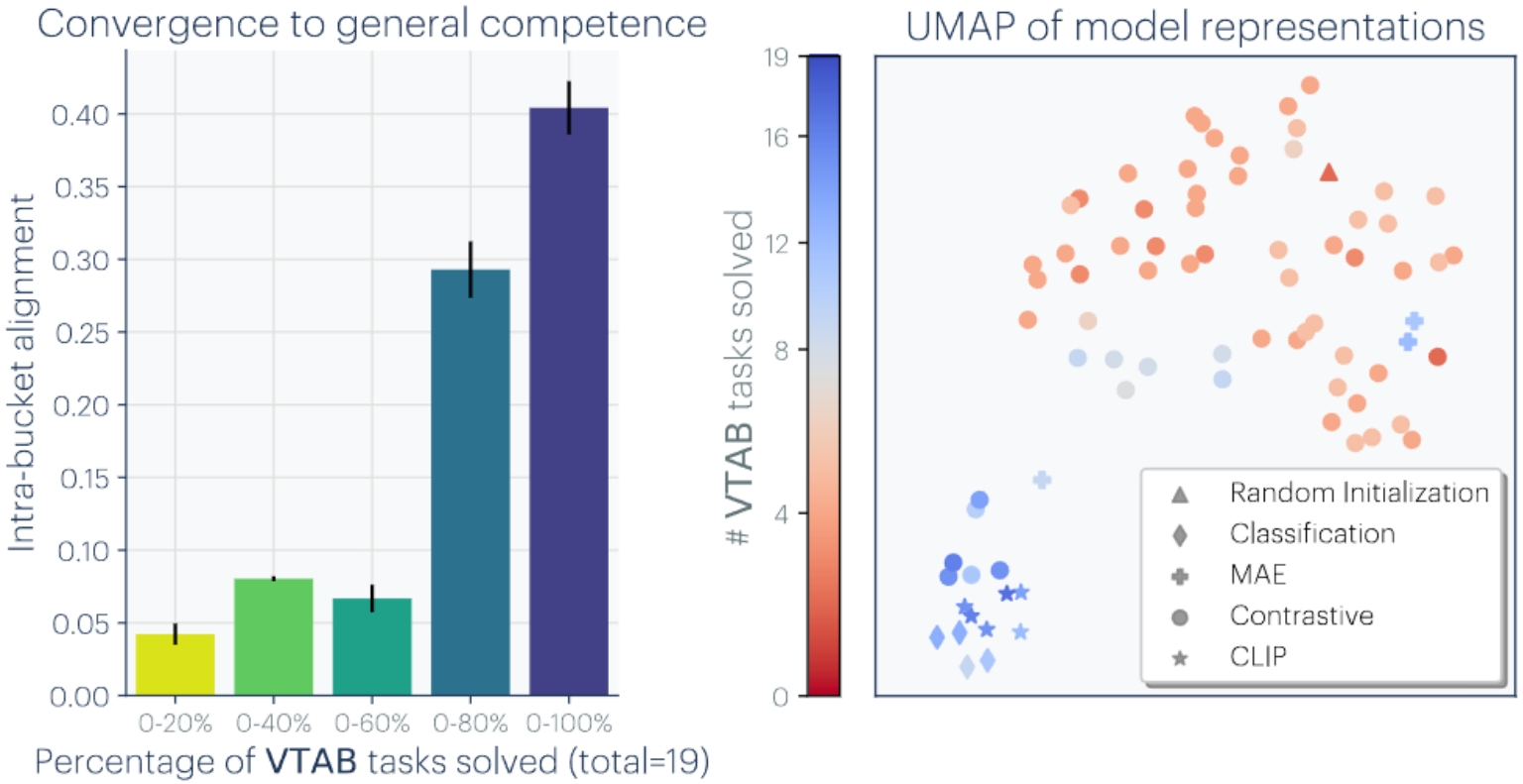


算法的训练数据是洞穴墙壁上的阴影，然而，我们假设，模型正在恢复洞穴外实际世界的更好表示。

核心观点 - 表征趋同，跨模态的趋同

表征趋同：不同的AI模型，不论其架构、训练目标或数据类型如何，其内部表示（表征）数据的方式正变得越来越相似。
跨模态的趋同：不仅相同类型的模型之间存在趋同，不同数据模态（如视觉和语言）的模型在表示数据时也显示出趋同的趋势。

随着时间的推移和跨多个领域，不同神经网络表示数据的方式变得越来越一致。然后，我们展示了跨数据模式的融合：随着视觉模型和语言模型变得越来越大，它们以越来越相似的方式测量数据点之间的距离：



强者大多相似，弱者各有不同:

左图：解决更多VTAB（视觉任务适应基准）任务的模型往往彼此更加一致。误差条显示标准误差。
右图：更有能力的和一般的模型（蓝色）有更多类似的表示。

为什么表征会趋同?

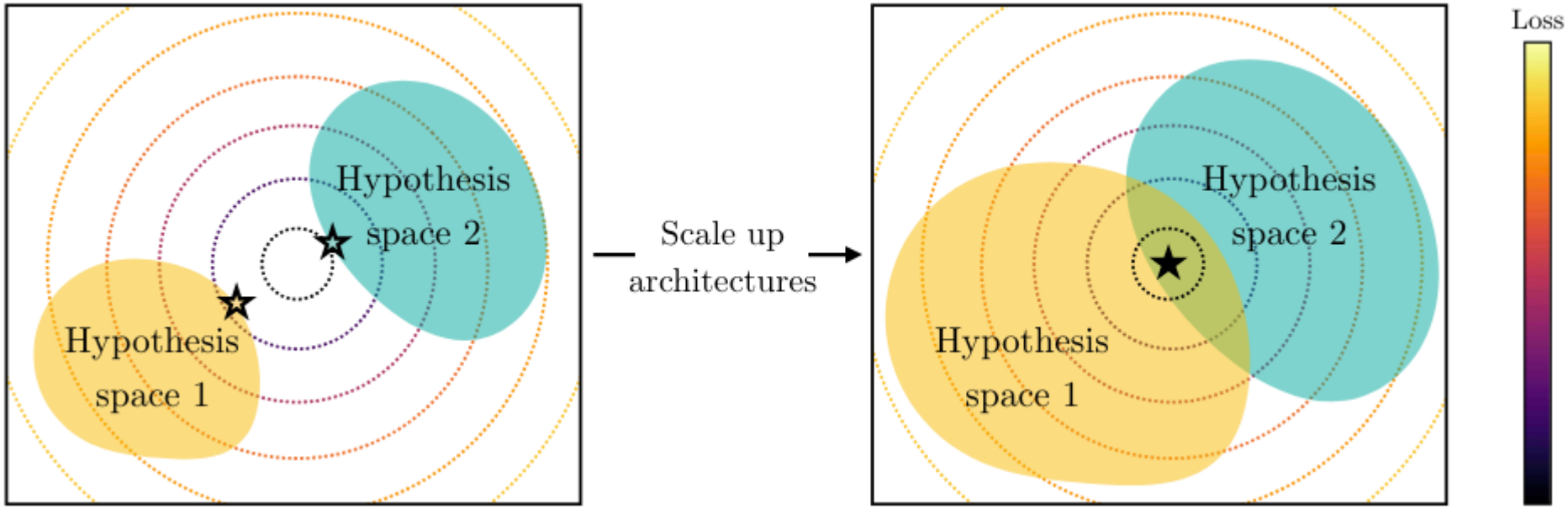
现代机器学习模型通常经过训练，以通过可能的隐式和/或显式正则化来最小化经验风险：

$$\overbrace{f^*}^{\text{trained model}} = \arg\min_{f \in \underbrace{\mathcal{F}}_{\text{function class}}} \mathbb{E}_{x \sim \text{dataset}} [\overbrace{\mathcal{L}}^{\text{training objective}}(f, x)] + \underbrace{\mathcal{R}}_{\text{regularization}}(f)$$

- **通过模型容量收敛**（紫色）：缩放模型（即，使用更大的函数类F）以及改进的优化，应该更有效地找到对该最优的更好的近似。
- **通过任务通用性进行收敛**（绿色）：能够胜任N个任务的表征比能够胜任M < N个任务的表征要少。当我们训练更多的通用模型来同时解决更多的任务时，我们应该期待更少的可能解决方案。
- **通过简单性偏差收敛**（红色）：深度网络偏向于寻找数据的简单拟合，模型越大，偏差就越大。因此，随着模型变大，我们应该期望收敛到更小的解空间。

为什么表征会趋同？- 1. 通过模型容量收敛

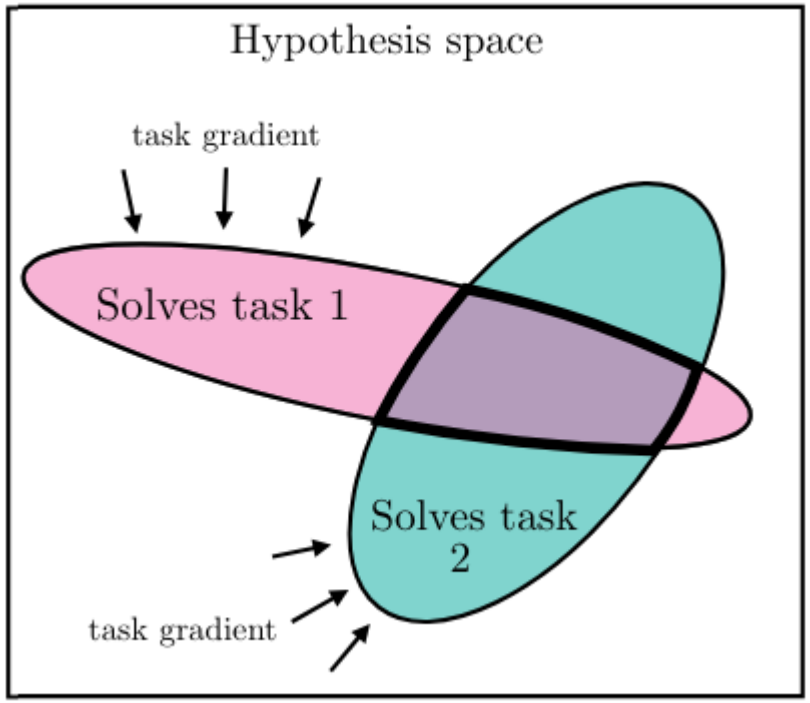
能力假设：
较大的模型比较小的模型更有可能收敛到共享表示。（**如果函数空间中存在一个最优表示，则更大的假设空间更有可能覆盖它。**）



左图：两个小模型可能不覆盖最优，从而找到不同的解决方案（标记为空心星号☆）。
右图：随着模型变得越来越大，它们覆盖了最优解并收敛到相同的解（实心星号）。

为什么表征会趋同？- 2. 通过任务通用性进行收敛

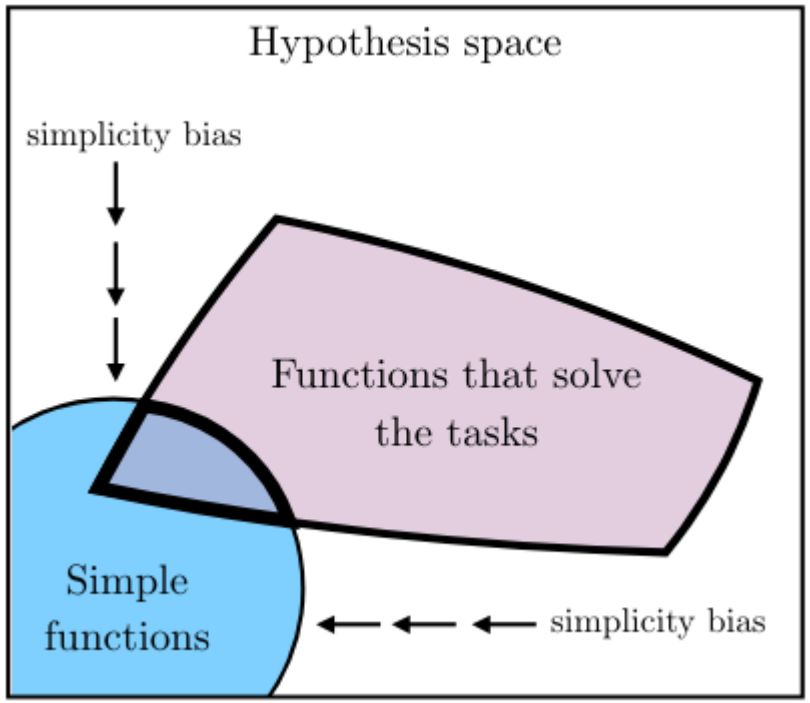
多任务尺度假设：
能够胜任N个任务的表征比能够胜任M < N个任务的表征要少。当我们训练更多的通用模型来同时解决更多的任务时，我们应该期待更少的可能解决方案。



每个训练数据点和目标（任务）都对模型施加了额外的约束。随着数据和任务规模的扩大，满足这些约束的表示量必须相应地变小。

为什么表征会趋同？- 3. 通过简单性偏差收敛

简单性偏见假说：
深度网络偏向于寻找数据的简单拟合，模型越大，偏差就越大。因此，随着模型变大，我们应该期望收敛到更小的解空间。



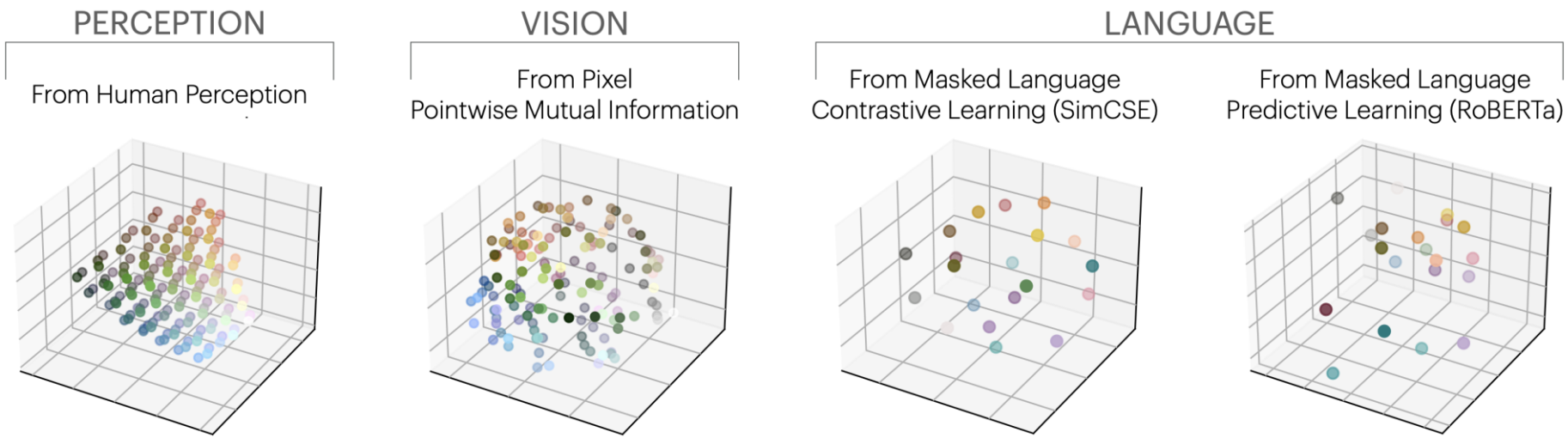
较大的模型具有更大的覆盖面的所有可能的方式来适应相同的数据。然而，深度网络隐含的简单性偏见鼓励更大的模型找到这些解决方案中最简单的。

向什么样的表征收敛?

在一个特定的理想化世界中，论文证明了一个特定的学习者家族将收敛到一个表示，其内核等于导致我们观察的潜在事件（Z）上的逐点互信息（PMI）函数，而不管模态如何。例如，在一个色彩世界中，事件 z_{red} 和 z_{orange} 会产生视觉和文本观察，我们会有：

$$\text{sim}(f(\texttt{"red"}), f(\texttt{"orange"})) = \text{PMI}(z\text{red}, z\text{orange}) + \text{const}$$

这种分析提出了各种假设，应该被视为更全面理论的起点。尽管如此，根据经验，论文确实发现像素颜色上的PMI恢复了与人类对颜色感知相似的内核，这也类似于LLMs恢复的内核：



点互信息PMI（Pointwise Mutual Information）

点互信息PMI（Pointwise Mutual Information）这个指标来衡量两个事物之间的相关性：

$$PMI(x;y) = log\frac{p(x,y)}{p(x)p(y)} = log\frac{p(x|y)}{p(x)} = log\frac{p(y|x)}{p(y)}$$

如果 x 跟 y 相互独立，则 $p(x,y) = p(x)p(y)$ 。二者相关性越大，则 $p(x,y)$ 就相比于 $p(x)p(y)$ 越大。

贝叶斯公式：

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

$$P(AB) = P(A)P(B|A)$$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

如何衡量表征是否收敛?

我们根据它们的内核来表征表示，即它们如何测量输入之间的距离/相似性。如果两个表示的核对于对应的输入是相同的，则认为它们是相同的。然后我们说这些表示是对齐的。

例如，如果文本编码器 $f$text 与图像编码器 $f$img 对齐，则我们将具有如下关系：

$$\text{sim}(f\text{text}(\text{"apple"}), f\text{text}(\text{"orange"})) \approx \text{sim}(f\text{img}(🍎), f\text{img}(🍊))$$

反例和局限性

- **双射投影假设**：论文中的数学论证仅严格适用于Z的双射投影，这意味着所有投影中的信息等同于底层世界的信息。这并不适用于有损或随机的观测函数。
- **目前并非所有表征都在汇聚**：虽然论文主要关注视觉和语言两种模态，但尚未观察到所有领域中的类似收敛现象。例如，在机器人领域，尚未形成标准化的方法来表示世界状态，与图像和文本表示的方式不同。
- **不同模态可能包含不同的信息**：不同模态（如视觉和语言）可能包含独特的信息，这些信息无法完全通过另一种模态来表达。例如，语言难以描述观看日全食的体验，而图像也无法传达“我相信言论自由”这样的抽象概念。
- **社会学偏见**：AI模型的发展轨迹受到研究者偏见和AI社区集体偏好的影响，这可能导致模型趋于模仿人类推理和表现，即使存在其他类型的智能。
- **特定用途智能可能不会汇聚**：不同的智能系统可能被设计来完成不同的任务，这些任务之间可能没有太多共同点。例如，生物信息学系统预测蛋白质结构，而自动驾驶车辆可能在高速公路上遵循车道。对于这些特定用途，可能存在更有效的表示方法，与现实世界的表示相脱离。

反例和局限性

- **如何测量对齐**：论文中使用了特定的对齐度量方法（如互近邻），但关于这些度量方法的优点和缺点存在争议。
- **需要进一步解释的现象**：尽管不同模型的表示趋于相似，但并非完全相同。例如，对齐度量分数可能只达到0.16，这表明仍有许多差异需要解释。
- **连续性和非有界世界**：论文中的理想化世界模型假设了一个离散的事件序列，但在现实世界中，我们可能需要处理连续的和非有界的观测。
- **随机观测**：论文中的模型假设了确定性的观测函数，但在现实世界中，观测可能是随机的，这可能影响模型的收敛性。
- **资源限制**：对于特定领域的持续扩展，如果遇到资源（如能源和计算能力）的限制，可能会影响模型的扩展和表现。


Thank
You