重新思考众包评估神经元解释

Tuomas Oikarinen
UC San Diego, CSE
toikarinen@ucsd.edu

Ge Yan
UC San Diego, CSE
geyan@ucsd.edu

Akshay Kulkarni
UC San Diego, CSE
a2kulkarni@ucsd.edu

Tsui-Wei Weng
UC San Diego, HDSI
lweng@ucsd.edu

Abstract

解释激活空间中的单个神经元或方向是机械解释性的重要组成部分。因此，已经提出了许多算法来自动生成神经元解释，但这些解释的可靠性常常不明确，或者不清楚哪些方法能产生最佳的解释。这可以通过众包评估来衡量，但它们常常噪声大且昂贵，导致结果不可靠。在本文中，我们仔细分析了评估流程，并开发了一种成本效益高且高度准确的众包评估策略。与先前的人类研究仅评估解释是否符合最强激活的输入不同，我们估计解释是否描述了所有输入中的神经元激活。为了有效地进行这种估计，我们引入了一种重要性抽样的新应用，以确定哪些输入最值得向评价者展示，导致与均匀抽样相比，成本减少约 30× 。我们还分析了众包评估中的标签噪声，并提出了一种贝叶斯方法来聚合多个评价，从而在保持相同准确度的情况下进一步减少所需评价数量 ∼ 5× 。最后，我们使用这些方法进行了一项大规模研究，比较两种不同视觉模型中最流行方法生成的神经元解释质量。

1 介绍

尽管它们具有变革性的能力，深度学习模型仍然本质上是不透明的，这限制了它们的可靠性和可信度，尤其是在高风险领域。为了解决这个问题，旨在理解神经网络内部机制的机械解释性领域迅速流行起来。机械解释性的一个关键部分是理解神经网络的小组件，例如稀疏自编码器（SAE）中的神经元或潜变量，这可以通过自动化神经元描述来实现 [3, 16, 11, 17, 18, 5, 14, 23, 1] 。

为了评估这些神经元解释的质量，研究人员通常依赖于人类研究，特别是在像 Amazon Mechanical Turk (AMT) 这样的平台上进行的众包评估。然而，尽管这些研究很重要，但其设计和分析却出乎意料地受到很少的关注。大多数评估依赖于主观的衡量标准，即解释是否与该神经元的最高激活输入相匹配。正如 [19] 所指出的，这只相当于测量召回率，并忽略了许多重要因素，例如描述是否描述了该神经元的较低激活，或者所有与解释匹配的输入是否实际激活了该神经元。

在本文中，我们的目标是使用一种更有原则的评估指标：相关系数，进行首次众包评估。然而，评估相关性给我们带来了新的挑战：

1. 成本和可扩展性。完美评估相关系数需要为数据集上的每一个输入收集概念的注释。估计一个（神经元，解释）对的相关性，使用三个独立的评级者大约需花费 $ 600 美元，使得在几百个神经元和多种解释方法之间的大规模比较变得不可能。这意味着我们需要从有注释的数据子集中有效地估计相关性。
2. 标签噪声和不确定性。AMT 等众包平台噪声很大。即使是很小的错误率也可能会破坏相关系数，尤其是在评估仅存在于少数输入中的稀有概念时，因为这时假阳性会比真阳性多。增加每个输入的评分人数可以帮助减少不确定性，但同时也会增加总体成本。

Preprint. Under review.

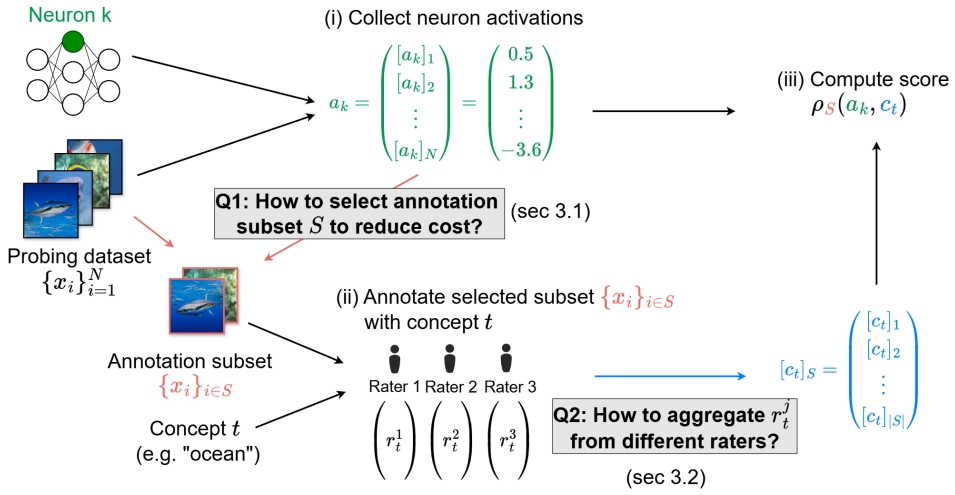


Figure 1: 针对神经元解释的众包评估流程。我们关注两个关键问题：Q1：“如何高效地选择用于标注的输入?” 和 Q2：“如何有效地处理标签噪声?”。我们提出的解决方案在第 3 节讨论，并在第 4 节验证。

在本文中，我们提出了新的方法来解决这些问题，并展示了我们的方法能够在保持准确性的同时，将评估相关系数的成本降低约 150× ，使我们的总评估成本从 $108,000 美元大幅下降到 $720 美元。我们的贡献如下：

- 我们进行第一次利用合适的评价指标，即相关系数，进行众包神经元解释评价
- 我们展示了如何有效利用重要性抽样来选择展示给评价者的最重要的输入，从而比均匀抽样减少 ∼ 30× 的标注成本。
- 我们开发了一种基于贝叶斯的方法来汇总不同评分者的预测，以处理噪声标签，进一步降低达到 ∼ 5× 精度的成本。
- 通过我们提出的方法，我们进行了一项大规模研究来比较不同的神经元解释方法，结果发现线性解释 [18] 整体上产生了最佳的视觉神经元解释，优于最近的基于 LLM 的解释 [1, 23] 及其他工作。

我们的代码将在 https://github.com/Trustworthy-ML-Lab/efficient_neuron_eval 上提供。

2 相关工作

2.1 视觉模型神经元的自动解释

基于标注概念数据的方法：或许第一个用于自动生成视觉模型中单个神经元文本描述的方法是网络解剖（Network Dissection, ND）[2] 。他们使用具有密集像素级标注的 Broden 数据集，并尝试找到与二值化神经元激活具有高交并比（IoU）的概念。然而，它仅限于使用具有密集概念标注数据的概念来搜索解释。组合解释（Compositional Explanations）[16] 扩展了网络解剖，以通过搜索诸如 “猫或狗” 这样逻辑组合的概念来处理可能激活多个不相关概念的多义神经元。聚类组合解释（Clustered Compositional Explanations, CCE）[14] 更进一步，解决了组合解释仅解释神经元最高激活的问题。他们将神经元的激活范围分为 5 个桶，并为每个桶生成一个单独的组合解释。[3] 提出一种利用分割模型而非标注数据的网络解剖版本。最后，[5] 提出用最大化 AUC 的逻辑组合数据集类（或超类）标签解释神经元，这消除了对 Broden 数据的依赖，但仍然需要标注数据。

基于生成语言模型的方法：另一种流行的方法涉及基于生成语言模型的描述。第一个这样的方法是 MILAN [11] ，它训练一个小型生成神经网络来描述一个神经元的最高激活输入。Describe-and-Dissect (DnD) [1] 是一种较新的方法，它利用预训练的语言模型而不是训练自己的模型，根据最高激活输入生成详细的神经元描述。多模态自动解释代理 (MAIA) 是另一种基于大型语言模型的解释管道，其中解释的 LLM 代理可以与多种工具交互，如查看高度激活的输入、生成新图像或编辑现有图像以生成其描述。

基于 CLIP 的方法：最后，一些论文如 CLIP-Dissect(CD) [17] 和 FALCON [13] 提出了无需依赖已标注概念信息的方法，而是依靠多模态模型的监督，如 CLIP [21] 。线性解释（LE）[18] 提出将多语义神经元解释为概念的线性组合，例如" $3 \times \text{dog} + 2 \times \text{cat}$ "。为了学习这些解释，他们要么利用数据集中的类别标签 - LE(label) - 要么使用来自 SigLIP [24] 的伪标签 - LE(SigLIP)。

2.2 关于神经元解释的人类研究

先前的工作 [2, 17, 1] 对神经元描述质量进行了众包评估。这些评估向评价者展示了导致该神经元高度激活的输入以及一个或多个描述，并要求用户评估这些描述与高度激活输入集的匹配程度。正如 [19] 指出的，这种评估方法存在缺陷，不能衡量描述是否匹配神经元的低激活，或所有与描述相符的输入是否都会导致神经元激活，并且实际上只测量了召回率。在本文中，我们主要关注解释与神经元激活之间的相关系数（跨所有输入）的评估，因为这是一个经过 [19] 提出的合理性检查的原则性评估指标。相关的，[25] 通过让评价者根据最高/最低激活输入来预测一个输入是否高度激活，对神经元的可解释性进行了大规模评估，但这项研究并没有使用任何神经元的文本描述。

3 方法

在本节中，我们将解决神经元解释的众包评估中的两个关键挑战：

- Q1: 如何在样本量较小的情况下高效估计解释 t 的质量?
- 问题二：我们如何在众包评估中有效处理标签噪声?

3.1 挑战 1：样本量小

问题表述。给定一个神经网络 f 和一个感兴趣的神经元 k ，可以手动或通过神经元解释方法自动获得神经元解释 t 。按照 [19] 介绍的框架，为了正确评估某个解释 t 是否描述了神经元 k 的行为，我们使用一个通过 [19] 描述的可靠性测试的指标，即皮尔逊相关系数，以测量神经元激活向量 a_k 与概念存在向量 c_t 之间的相似性。这里，$a_k \in \mathbb{R}^{|\mathcal{D}|}$ 是神经元 k 对所有探测图像的激活向量，$c_t \in \mathbb{R}^{|\mathcal{D}|}$ 是概念向量，用于指示概念 t 是否出现在探测图像上。i 的第 $[a_k]_i$ 和 $[c_t]_i$ 组件可以表示为：$[a_k]_i = f_k^{0:l}(x_i), \quad [c_t]_i = \mathbb{P}(t|x_i), \ \forall\, i \in \{|\mathcal{D}|\}$ 其中 $f^{0:l}$ 表示直到第 l 层的神经网络。

我们的目标是评估皮尔逊相关系数 ρ ，其可以表示为：

$$\rho(a_k, c_t) = \frac{1}{|\mathcal{D}|} \frac{\sum_{i \in \mathcal{D}} ([a_k]_i - \mu(a_k)) \cdot ([c_t]_i - \mu(c_t))}{\sigma(a_k)\sigma(c_t)} \tag{1}$$

其中 $\mu(a_k), \sigma(a_k), \mu(c_t), \sigma(c_t)$ 是向量 a_k 和 c_t 的均值和标准差。

激活向量 a_k 可以通过神经网络的前向传播轻松评估。然而，概念向量 c_t 可能会很昂贵且难以获取，例如需要众包标注。我们的问题变成了：我们如何能最佳选择一个输入子集 $S \subseteq \{|D|\}$ 来标注相应的概念标签 $\{[c_t]_i\}_{i \in S}$ ，以便我们可以准确估计 $\rho(a_k, c_t)$ 。

一个自然选择 S 的方法是使用蒙特卡罗（均匀）抽样，这种方式是从 D 中随机均匀选择一批样本 $\{x_i\}_{i \in S}$ 。注意，方程 (1) 可以写成期望的形式：

$$\rho = \mathbb{E}_{x_i \sim \mathcal{P}} \frac{([a_k]_i - \mu(a_k)) \cdot ([c_t]_i - \mu(c_t))}{\sigma(a_k)\sigma(c_t)}, \tag{2}$$

，这意味着我们可以从均匀分布 $\mathcal{P}$ 中抽样 x_i ，其中所有输入的 $p(x_i) = \frac{1}{|D|}$ 。随后我们可以将相关性视为测量 $\bar{a}_{ki} \cdot \bar{c}_{ti}$ 的期望值，在此其中（$\bar{a}_{ki} = ([a_k]_i - \mu(c_t))/\sigma(a_k)$ ，$\bar{c}_{ti} = ([c_t]_i - \mu(c_t)))/\sigma(a_k)$ 。通过蒙特卡罗抽样，估计值将是 $\rho_S = \frac{1}{|S|} \frac{\sum_{i \in S} ([a_k]_i - \mu(a_k)) \cdot ([c_t]_i - \mu(c_t))}{\sigma(a_k)\sigma(c_t)}$ ，其中 $\mu(c_t)$ 和 $\sigma(c_t)$ 是从样本中估计的。然而，这种方法产生一个问题：与探测数据 $|\mathcal{D}|$ 的规模相比，大多数概念 t 是罕见的，所以从 $\mathcal{P}$ 中取的小样本 $S = \{i_1, i_2, ..., i_{|S|}\}$ 可能会使 $[c_t]_{i \in S}$ 全部为 0——这意味着我们没有抽取任何包含概念 t 的输入，从而无法准确估计 ρ 。这引出了我们的第二个想法，重要性抽样。

为了解决这个问题，我们建议使用重要性采样来估计公式 (2) 中的 ρ ，也就是说，使用 $x_i \sim \mathcal{Q}$ ，其中 $\mathcal{Q}$ 是比原始均匀分布 $\mathcal{P}$ 更有利的概率分布。重要的是，重要性采样不会改变我们所

评价的函数的期望值，也就是说对于任何分布 Q ，均有 $\mathbb{E}_{x\sim\mathcal{P}}[h(x)] = \mathbb{E}_{x\sim\mathcal{Q}}[\frac{p(x)}{q(x)}h(x)]$ 。根据 [22] （第 3.3.2 节，定理 3.12），在估计函数 h 的期望值时，也就是 $\mathbb{E}_{x\sim\mathcal{P}}[h(x)]$ ，使估计器方差最小的最优采样分布 q^* 应该具有概率密度函数（pdf）$q^*(x) \propto |h(x)|p(x)$ 。正式的陈述和证明见附录 B.3 。

基于定理 1 和公式 (2) ，我们可以通过从 $\mathcal{Q}$ 中以概率密度函数 $q(x)$ 进行采样来估计 ρ ：

$$q(x_i) \propto |h(x_i)| \cdot \frac{1}{|\mathcal{D}|}, \quad h(x_i) = \frac{([a_k]_i - \mu(a_k)) \cdot ([c_t]_i - \mu(c_t))}{\sigma(a_k)\sigma(c_t)}. \tag{3}$$

。换句话说，为了最小化估计量 ρ 的方差，我们应该更多地从具有高乘积值 $\bar{a}_{ki} \cdot \bar{c}_{ti}$ 的探测图像中采样 x_i ——意思是说那些具有高神经元激活 $\bar{a}_{ki}$ 并包含概念 t 的图像。

由于在进行测试（例如人体研究或昂贵的模拟流程）之前我们并不知道 c_t ，我们使用一种更便宜（但不太精确）的方法来近似 c_t 以用于采样。具体来说，我们使用 SigLIP [24] 来预测 c_t^{siglip} 。因此，我们最终的采样分布 q 是：

在 $[\bar{c}_t^{\text{siglip}}]_i = \frac{[c_t^{\text{siglip}}]_i - \mu(c_t^{\text{siglip}})}{\sigma(c_t^{\text{siglip}})}$ 和 $\epsilon = 0.001$ 的情况下，必须确保所有输入的收敛和非零抽样概率。

采样校正。为了使用重要性采样对 $\mu(c_t)$ 和 $\sigma(c_t)$ 进行无偏估计，我们需要依次应用重要性采样校正。首先，我们估计均值 $\mu_S(c_t)$ ，然后从样本 $x_i \sim \mathcal{Q}$ 估计标准差 $\sigma_S(c_t)$ ：

$$\mu_S(c_t) = \frac{1}{|S|}\sum_{i\in S}\frac{p(x_i)}{q(x_i)}[c_t]_i, \quad \sigma_S(c_t) = \sqrt{\frac{1}{|S|-1}\sum_{i\in S}\frac{p(x_i)}{q(x_i)}([c_t]_i - \mu_S(c_t))^2}. \tag{4}$$

接下来，我们对输入进行标准化：$[\bar{c}_t]_S = \frac{[c_t]_S - \hat{\mu}_S(c_t)}{\hat{\sigma}_S(c_t)}$ ，$[\bar{a}_k]_S = \frac{[a_k]_S - \mu_S(a_k)}{\sigma_S(a_k)}$ ，并最终获得 ρ_S 的估计相关评分：

$$\rho_S = \frac{1}{|S|}\sum_{i\in S}\frac{p(x_i)}{q(x_i)}[\bar{a}_k]_i \cdot [\bar{c}_t]_i \tag{5}$$

我们面临的另一个挑战是，通常我们得到的概念标签是嘈杂的。这对于诸如亚马逊机械土耳其人之类的众包人类评估尤其是一个问题。为了解决这个问题，我们可以为每个输入收集二进制评分，并使用不同的方法对它们进行汇总。设为特定的（输入，概念）对的评分集。下面我们描述了三种不同的评分汇总方法，即汇总标注以获得最终标签：

方法 1 和方法 2 是汇总来自多个来源的标签的常用技术，因为它们非常简单且直观。然而，正如我们在第 4.2 节中所展示的，我们可以通过利用贝叶斯法则来估计 $[c_t]_i$ 为 $\mathbb{P}([c_t^*]_i = 1|R_{ti})$ ，从而进一步改进。在这里，我们使用 c_{ti}^* 表示没有噪音的“真实”概念值。

- 方法三 - 贝叶斯：

$$[c_t]_i = \frac{\mathbb{P}(R_{ti}|c_{ti}^* = 1) \cdot \mathbb{P}(c_{ti}^* = 1)}{\mathbb{P}(R_{ti}|[c_{ti}^* = 1) \cdot \mathbb{P}(c_{ti}^* = 1) + \mathbb{P}(R_{ti}|c_{ti}^* = 0)(1 - \mathbb{P}(c_{ti}^* = 1))} \tag{6}$$

我们使用贝叶斯定理将后验概率 $\mathbb{P}([c_t]_i = 1|R_{ti})$ 按方程 (6) 进行展开，并按如下方法计算每一项：

(I) 似然度：$\mathbb{P}(R_{ti}|c_{ti}^*)$ 。假设每个评估者均以误差率 η 随机地犯错误，即对于任何输入 $\mathbb{P}(r_{ti}^j = c_{ti}^*) = 1 - \eta$ ，其中 η 是一个可以通过实验估计的参数。令 $\alpha_{ti} = \sum_{j=1}^m r_{ti}^j$ ，我们得到如下方程中的似然度：

$$\mathbb{P}(R_{ti}|c_{ti}^* = 1) = (1-\eta)^{\alpha_{ti}}(\eta)^{(m-\alpha_{ti})}, \quad \mathbb{P}(R_{ti}|c_{ti}^* = 0) = (\eta)^{\alpha_{ti}}(1-\eta)^{(m-\alpha_{ti})} \tag{7}$$

(II) 先验: $\mathbb{P}(c_{ti}^{gt})$ 。有多种选择先验的方法，反映了输入 x_i 中是否存在某个概念的信心。在我们的分析中，我们考虑了两种不同的先验：

- (a) 均匀先验：对于所有概念 t 和所有输入 x_i ，设定 $\mathbb{P}(c_{ti}^* = 1) = \beta$
- (b) SigLIP 先验：我们利用来自廉价评估器的知识，即 SigLIP 来初始化先验并设置 $\mathbb{P}(c_{ti}^* = 1) = [c_t^{siglip}]_i$

这里 β 是一个超参数，我们将其设置为 $\beta = 0.01$ 。对于 SigLIP 先验，我们将先验限制在 0.001 和 0.999 之间，以避免可能主导最终结果的极端值。使用 SigLIP 先验的方法可以看作是一种结合了人类和模型知识的混合评价。

4 方法测试与验证

在本节中，我们设计了两个设置来测试我们在第 3 节中提出的解决方案是否能有效处理神经元解释中众包评估的小样本和噪声标注：设置 1 利用带有人工标签噪声的模拟，设置 2 进行真实的小规模众包研究。本节提供了定量结果和关于哪些方法可以作为在第 5 节中进行的大规模众包研究的成本有效解决方案的指导。

场景 1：模拟人类研究

在这种设置中，我们进行模拟来研究当我们使用数据集中探测图像 x_i 的真实类别（或超类）标签用于 c_t 时估计量 ρ_S 的性能。在这种情况下，c_t 被标记为 c_t^{gt} ，以区分直接从 AMT 获取的概念标签（在设置 2 中会更加嘈杂）。首先，我们通过选择来自 ImageNet 类和超类标签的、与神经元 k 的激活相关性最大的概念 t_k ，来描述在 ImageNet 上训练的 ResNet-50 [10] 的 layer4 中的神经元，使用完整探测数据集 $\mathcal{D}$ 中的真实概念标签 c_t^{gt} ：

$$t_k = \text{argmax}_{t \in \mathcal{C}} \ \rho(a_k, c_t^{gt}) \tag{8}$$

这个概念在整个数据集上与正确标签的相关系数作为我们的真实相关系数：

$$\rho_{gt} = \rho(a_k, c_{t_k}^{gt}). \tag{9}$$

在这个设置中，我们可以通过估计探测数据 x_i 输入子集上的相关性来模拟一个 Mechanical Turk 研究，$i \in S, S \subset \mathcal{D}$ ：

$$\rho_S = \rho([a_k]_S, [c_t^{gt}]_S) \tag{10}$$

为了模拟挑战 2（第 **??** 章）中 c_t 的标签噪声，我们随机翻转了一定比例的 c_t^{gt} 到相反的标签。这里我们使用的噪声率为 η =13 %，这是我们在 Mechanical Turk 上测量到的粗略错误率。对于每次评估（即图 2a 中的每个数据点），我们在 10 次不同的试验中平均了估计误差。

设置 2：小规模众包研究在这种设置下，我们在 AMT 上进行了真正的众包评估，以为探测数据集中的样本 x_i 注释概念标签 c_t ，而不是使用设置 1 中的数据集的真实标签 c_t^{gt} 。

具体而言，我们对 resnet-50 中的 5 个神经元进行了众包评估，解释则从 ImageNet 类和超级类中根据方程 (8) 选出，并估计相关系数为

$$\rho_S = \rho([a_k]_S, [c_t^{amt}]_S). \tag{11}$$

。这种设置使我们可以通过测量在方程 (11) 中估计的相关系数与方程 (9) 中“真实值”相关系数的差异，来评估 Mechanical Turk 上的错误率和不同抽样策略的有效性。在这种情况下，我们选择了具有清晰明确概念的神经元，例如蜥蜴，并由 9 位 AMT 评价者每个输入标注 300 个输入。为了估计较少数量的评价者/输入的评估准确性，我们随机抽取评分子集，用于图 2b 和 3 中的绘图（注意：每个点是相同大小子集 1000 个随机样本的平均值）。我们还使用此设置估计 MTurk 评价者的错误率为 $\eta = \mathbb{P}(r_{ti}^j \neq [c_t^{gt}]_i)$ ，给出了 13 % 的估计值。评估指标：我们比较不同抽样策略的主要评估指标是在设置 1 和 2 中，真实相关性 ρ_{gt} 和估计相关性 ρ_S 之间的相对相关误差（RCE）：

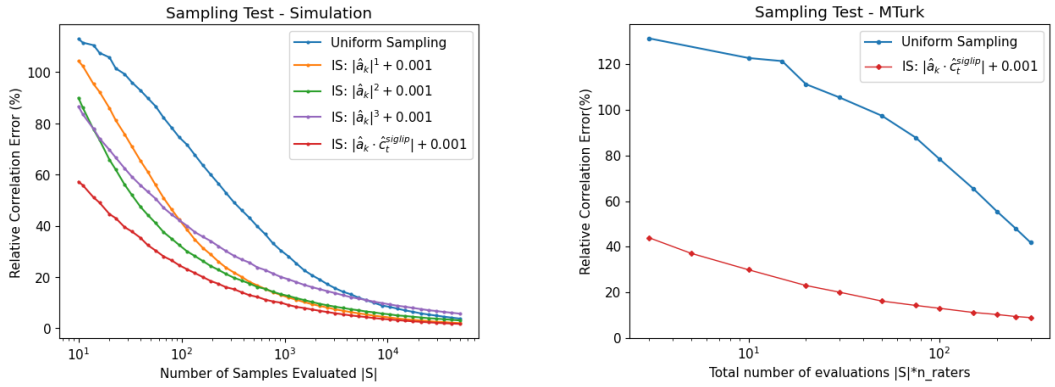
$$RCE = \frac{1}{K} \sum_{k \in K} \left| \frac{\rho_S(a_k, c_{t_k}) - \rho_{gt}(a_k, c_{t_k}^{gt})}{\rho_{gt}(a_k, c_{t_k}^{gt})} \right| \tag{12}$$

例如，RCE 为 20 % 表示平均而言，我们估计的相关性比真实的相关性值偏差 20 %。

4.1 结果：采样方法

图 2 显示了在我们的两个设置中不同采样策略的结果。我们可以看到，总体而言，在方程 (**??**) 中定义的重要性采样表现最好（红线），因为左下角代表低误差和低成本。与均匀采样（蓝线）相比，我们可以用大约 30× 更少的样本达到相似的相关误差，从而在标注成本上减少 30× 。这对于我们的无标签噪声模拟结果（图 2a ）以及在 MTurk 测试中使用真实评价者的情况（图 2b ）都是成立的。

除了方程 **??** 中定义的均匀采样和最佳重要性采样之外，我们还测试了仅依赖于神经元激活 a_k 的重要性采样。虽然其表现不如我们的 SigLIP 指导采样，但这种采样比均匀采样提供了显著的改进，并且基于我们的实验，若没有廉价的 c_t 估计器可用，使用 $q(i) \propto [\hat{a}_k]_i^2 + \epsilon$ 可以是一个不错的替代选择。基于这些分析，我们选择在第 5 节的大规模众包评估中使用重要性采样（方程 **??** ）。



(a) 情境 1 中的模拟结果

(b) 设置 2 中的 MTurk 测试结果

Figure 2: 在章节 4 中比较两种环境下的不同采样策略。我们可以看到，使用 SigLIP 估计的重要性采样（红色）明显优于其他方法。

4.2 结果：评分聚合方法

在这里，我们评估了在第 **??** 节中提出的不同评价聚合策略，以解决标签噪声的问题。正如图 **??** 所示，方法 3b（贝叶斯方法与 SigLIP 先验，红线）在两个环境下的整体 RCE 最低，相比其他方法在约 5× 降低成本的情况下达到了类似的错误率。方法 3a（贝叶斯方法与均匀先验，绿线）和方法 2（多数投票，橙线）的表现相似，方法 3a 在评估次数较多时略胜一筹。在图 **??** 中，每条线代表聚合方法通过使用给定成本的最佳评估人数可以达到的最低错误率。在这些分析的基础上，我们选择使用方法 3b：贝叶斯方法与 SigLIP 先验来汇总我们在第 5 节中的大规模众包评估数据。需要注意的是，我们还报告了使用方法 3 与均匀先验的结果，以展示纯粹的人类评估结果。

我们也可以利用这些实验来确定我们人类研究的最佳参数，具体讨论详见附录 A.2 。基于我们的发现和预算，我们选择对每个输入使用 2 名评分员，并评估每个神经元的 90 个输入。由于我们可以在一个任务中以 $ 0.06 的成本评估 15 个输入，这项评估将花费 $\frac{\$0.06}{15} \cdot 90 \cdot 2 = \$0.72$ 每（神经元，解释）对。

5 大规模众包研究

设置。我们的研究重点是在两个不同模型的两个层级上的神经元：在 ImageNet 上训练的 ResNet-50 的 Layer4，以及在 ImageNet 上训练的 ViT-B-16 [6] 的 Layer11 中的 MLP 激活。我们使用了完整的 ImageNet 验证集作为探测数据集，详见附录 B.1 。

5.1 通过自动化评估选择研究方法

已经有许多不同的方法被引入用于解释视觉模型中的神经元。由于预算有限，我们无法在众包研究中比较所有这些方法，因此我们首先进行自动评估以找出在众包研究中包括的最佳方法。对于自动评估，我们使用由 [18] 引入的基于 SigLIP 的模拟与相关评分。

解释复杂度：一些方法把单个神经元解释为概念的组合。例如，[16, 5] 使用逻辑组合的解释，如 “狗或猫”，而 [18] 提议将神经元描述为概念的线性组合，如 “$2.7 \cdot \text{dog} + 1.5 \cdot \text{cat}$ ”。更复杂的解释通常更精确，但更难理解，并且评估成本更高，因为它需要解释中涉及的每个概念的标签。我们使用长度 l 来表示解释复杂度，其中 l 是解释中唯一概念的数量。

为了公平起见，我们将解释分为两个组，简单解释是 $l = 1$ ，复杂解释是 $l > 1$ 。表格 1 展示了不同简单解释的比较，而表格 2 比较了复杂的解释方法。总体而言，我们可以看到线性解释 [18] 产生了最高的相关性评分。其他表现良好的方法包括 INVERT [5] 和 CLIP-Dissect [17] ，随后是最近基于语言模型的解释 MAIA [23] 和 DnD [1] 。我们观察到基于 Broden 的方法 [2, 16, 14] 整体表现相对较差，可能是因为它们的概念集不包括描述 ImageNet 模型后期神经元所需的相关高级概念，如动物种类。

Simple Exp. (l =1)	ND [3]	MILAN [11]	CD [17]	INVERT l=1 [5]	DnD [1]	MAIA [23]	LE(label) l=1 [18]	LE(SigLIP) l=1 [18]
RN-50 (Layer4)	0.1242 ± 0.002	0.0920 ± 0.002	0.1904 ± 0.002	0.1867 ± 0.002	0.1534 ± 0.002	0.1396 ± 0.009	0.1793 ± 0.002	0.2413 ± 0.002
ViT-B-16 (Layer11 MLP)	0.0335 ± 0.002	0.0194 ± 0.002	0.1849 ± 0.004	0.1343 ± 0.004	0.1049 ± 0.003	0.1497 ± 0.021	0.2704 ± 0.004	0.2968 ± 0.004

Table 1: 基于 SigLIP 的模拟通过相关性评分比较不同的简单解释方法（$l = 1$ ）。我们可以看到，即使在长度为 1 的情况下，线性解释（LE）表现得最好。

Complex Exp. (l >1)	Comp Exp [16] l=3	INVERT [5] l=3	CCE [14], l= 5	CCE [14], l=15	LE(label) [18] l=4.37/1.97	LE(SigLIP) [18] l=4.66/1.82
RN-50 (Layer4)	0.1399 ± 0.002	0.2341 ± 0.002	0.0993 ± 0.002	0.1510 ± 0.003	0.2924 ± 0.002	0.3772 ± 0.002
ViT-B-16 (Layer11 MLP)	0.0468 ± 0.002	0.1101 ± 0.003	0.0534 ± 0.003	0.0570 ± 0.006	0.3243 ± 0.005	0.3489 ± 0.005

Table 2: 基于 SigLIP 的模拟使用相关评分来比较复杂解释（$l > 1$ ）。我们可以看到线性解释（LE）的整体表现最好。

5.2 众包评估

设置。基于我们在前一章节的研究结果，我们选择了对 5 个表现最好的简单解释进行众包研究，分别是 LE(SigLIP) [18]，CLIP-Dissect [17]，INVERT [5]，MAIA [23] 和 DnD [1]。我们为每个模型随机选择了 100 个神经元进行评估，并为每个输入由 2 位评分员进行评价，每个神经元 90 个输入。总共评估了 1000 对 (神经元, 解释) 组合，总成本为 $ 720。根据我们机构的 IRB 审核委员会的决定，我们的研究被认为无需 IRB 批准。

用户界面。我们众包研究的主要目标是让用户标注哪些输入中存在概念 t 。我们通过向评估者展示解释 t 和每个任务 15 个输入，来请他们选择概念存在的输入。完整的研究界面如图 7 所示。

5.2.1 结果

Model	Aggregation	CD [17]	INVERT l=1 [5]	DnD [1]	MAIA [23]	LE(SigLIP) l=1 [18]
RN50 (Layer4)	Bayes (Uniform Prior)	0.1500 ± 0.013	0.1258 ± 0.009	0.1576 ± 0.013	0.1215 ± 0.013	0.1783 ± 0.011
	Bayes (SigLIP Prior)	0.1817 ± 0.010	0.1637 ± 0.006	0.1661 ± 0.011	0.1309 ± 0.009	0.2107 ± 0.010
ViT-B-16 (Layer11 mlp)	Bayes (Uniform Prior)	0.1230 ± 0.016	0.0770 ± 0.016	0.0670 ± 0.015	0.0938 ± 0.017	0.1513 ± 0.018
	Bayes (SigLIP Prior)	0.1860 ± 0.020	0.1210 ± 0.021	0.0895 ± 0.017	0.1516 ± 0.021	0.2382 ± 0.022
Avg:		0.1601	0.1219	0.1201	0.1245	0.1946

Table 3: 我们大规模 MTurk 评估的结果。每一行代表了 100 个随机神经元的平均相关性，以及均值的标准误差。

众包评估的总体结果如表 3 所示。在图 5 中，我们展示了一些示例描述及其评分。可以看到，线性解释尽管局限于使用预定义的概念集，但在整体表现上明显最佳。我们认为这是因为它是唯一一个旨在解释整个激活范围的方法，而不仅仅关注最高激活。第二个表现最好的方法是 CLIP-Dissect [17]，而其他方法的表现基本相当，具体表现因所评估的模型而有所不同。有些令人惊讶的是，基于大型语言模型 [1, 23] 的最新方法并没有超越较为简单的基准，尽管它能够在某些情况下提供非常准确和复杂的描述。我们认为这有几个原因：

1. 仅关注高激活输入，导致过于具体的解释，这些解释不能描述较低的激活。

2. 不一致性。虽然所有 5 种方法通常产生相关的描述，但更复杂的方法（例如那些依赖于 LLM 的方法）在描述质量上往往具有更大的差异，导致对某些神经元的解释较差。

最后，总体上我们发现那些在自动化 SigLIP 评分中表现较好的方法 (表 1) 在人工研究中 (表 3) 也表现良好。虽然相关性评分整体下降了（可能由于标签噪声），方法之间的顺序保持稳定，突出了基于 SigLIP 的解释评估的可靠性。

在本文中，我们引入了新的方法来对神经元解释进行有原则的众包评估。首先，我们提出使用重要性采样来决定向评价者展示哪些输入，从而将我们的标注成本降低 $\sim 30\times$ 。其次，我们介绍了一种基于贝叶斯法则的新方法来聚合多个评价，这进一步减少了在达到一定准确度水平时所需的评价数量，又减少了 $\sim 5\times$ 。最后，使用这些方法，我们能够在合理的成本下（将成本从估计的 $ 108,000 美元降低到 $ 720 美元）进行大规模的人体研究，而不牺牲准确性——我们评估了 5 种表现最佳的视觉神经元解释方法，发现线性解释 [18] 的表现最好，超过其他方法。

References

[1] Nicholas Bai, Rahul A Iyer, Tuomas Oikarinen, Akshay Kulkarni, and Tsui-Wei Weng. Interpreting neurons in deep vision networks with language models. *TMLR*, 2025.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *PNAS*, 2020.

[4] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.

[5] Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina MC Höhne. Labeling neural representations with inverse recognition. In *NeurIPS*, 2023.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[7] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip. *arXiv preprint arXiv:2406.04341*, 2024.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[9] Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. Enhancing automated interpretability with output-centric feature descriptions. *arXiv preprint arXiv:2501.08319*, 2025.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *ICLR*, 2022.

[12] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[13] Neha Kalibhat, Shweta Bhardwaj, C Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *ICML*, 2023.

[14] Biagio La Rosa, Leilani Gilpin, and Roberto Capobianco. Towards a fuller understanding of neurons with clustered compositional explanations. *Advances in Neural Information Processing Systems*, 36, 2023.

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

[16] Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In *NeurIPS*, 2020.

[17] Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. In *ICLR*, 2023.

[18] Tuomas Oikarinen and Tsui-Wei Weng. Linear explanations for individual neurons. In *International Conference on Machine Learning*, 2024.

[19] Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. In *International Conference on Machine Learning*, 2025.

[20] OpenAI. GPT-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2024.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[22] Christian Robert and George Casella. *Monte Carlo Statistical Methods, second edition*. Springer New York, 2004.

[23] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

[24] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

[25] Roland S. Zimmermann, Thomas Klein, and Wieland Brendel. Scale alone does not improve mechanistic interpretability in vision models. In *NeurIPS*, 2023.

A 讨论

由于时间和预算限制，我们的众包评估集中在比较几个模型/层的描述，例如 ImageNet 训练模型的后层神经元。虽然我们大部分发现 ResNet 和 ViT 模型之间的趋势相似，不同的描述方法可能在不同类型的神经元上有优势。例如，如果我们专注于较低层神经元或训练在 Places365 上的模型，基于网络解析的方法可能会比在当前评估中表现得更好，因为 Broden 数据集中的标签更适合这些任务。同样，我们认为基于 LLM 的方法在描述稀疏自编码器的神经元时可能表现更好，因为这些神经元更具单义性，仅通过高激活输入就能更好地描述。

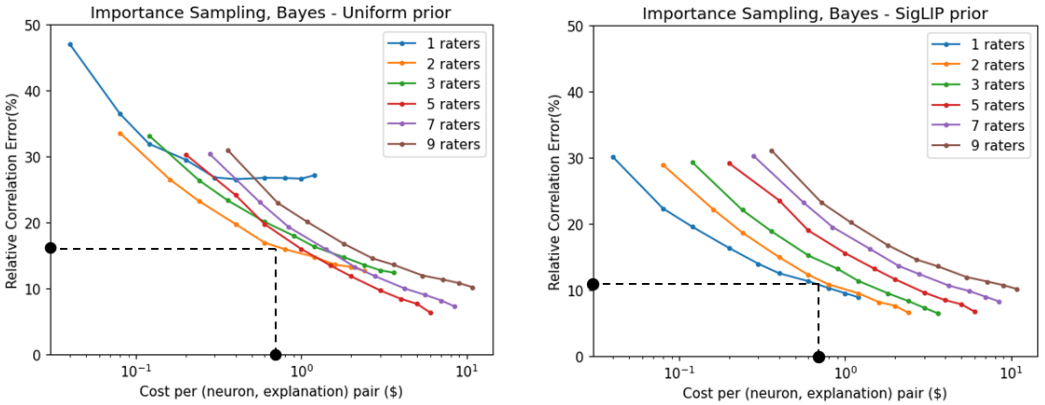
其次，我们的众包评估依赖于 Amazon Mechanical Turk 的工人，他们不是专家，常常在标记时出错。虽然我们引入了原则性措施来估计错误并缓解错误，但我们无法提升他们的领域知识，这意味着众包评估可能偏向于较简单的描述，而不是需要领域知识的更复杂的概念。要比较更复杂的描述或特定领域的神经元描述，可能需要招募领域专家来进行评估。

最后，我们的评估侧重于评估基于输入的神经元解释，这些解释旨在解释“输入 → 神经元激活”函数。而一些最近的工作如 [7, 9] 则专注于基于输出的神经元解释，这些解释旨在解释“神经元激活 → 模型输出”关系。严格评估这些基于输出的解释将需要不同的方法学，并且是未来工作的一个有趣问题。

A.1 更广泛的影响

本文旨在通过可解释性更好地理解神经网络，因此我们预计其影响主要是积极的，因为更好地理解神经网络可以帮助我们在部署之前识别故障模式，并实现对模型更好的控制。由于我们的重点尤其放在严格评估神经元解释上，这可以帮助避免可解释性的错觉或用户过度依赖不可靠的解释。

A.2 众包研究设计



(a) 方法 3 的聚合：贝叶斯 - 均匀先验。 (b) 聚合方法 3：贝叶斯 - SigLIP 先验。

Figure 3: 在第 4 节描述的设定 2（在 MTurk 上测试）中比较评级聚合策略。x 轴表示评估的成本，即每个神经元的输入数量乘以每个输入的评价者数量再乘以每个输入的成本。最低的曲线代表特定预算的最佳评价者数量。

在本节中，我们讨论如何利用第 4 节中描述的方法测试/验证设置来帮助设计研究的更细微参数，比如每个输入使用多少评审员。在我们当前的 MTurk 设置中，每个单一评审员对单张图像进行评分的成本为 $\frac{\$0.06}{15} = \$0.004$，因为我们在每个任务中展示 15 个输入。然后，我们可以通过在 x 轴上绘制 $n_{inputs} \cdot n_{raters} \cdot \$0.004$ 来将预期误差作为评估成本的函数进行绘制。

在图 3 中，我们绘制了每个输入有不同评分者数量时的成本与预期错误率。在大多数评估预算情况下，使用 2 个评分者时，贝叶斯方法配合均匀先验能够给出最佳结果，而使用 1 个评分者时，贝叶斯方法配合 SigLIP 先验能够给出最佳结果。根据我们的预算，我们的目标是每个（神经元，解释对）大约 $ 0.7 的成本，我们可以看到在我们的图中顺着黑线寻找该预算的最佳评分者数量和预期错误率。由于我们希望在使用两种聚合方法时都能获得良好的

结果，并且在这个范围内 SigLIP 采样无论是 1 个还是 2 个评分者几乎给出相同的错误，我们选择在众包研究中使用 2 个评分者和每个神经元 90 个输入。这使我们根据 MTurk 测试达到关于贝叶斯 - 均匀先验的预期相关错误约为 16%，关于贝叶斯 - SigLIP 先验的预期相关错误约为 12%。

B 附加细节

B.1 实验细节

我们的评估集中在两个模型/层上：

1. 在 ImageNet 上训练的 ResNet-50 的 Layer4（残差块 4 的末端）。对于模拟，我们使用在平均池化后的神经元激活，给予标量激活，但对于为二维激活设计的方法，平均池化前的激活作为输入。
2. 在 ImageNet 上训练的 ViT-B-16 [6] 编码器的第 11 层的 MLP 激活。我们只关注 CLS-token 的激活，因为这是最后一层，其他 token 不影响预测。

我们使用完整的 ImageNet 验证数据集作为 $\mathcal{D}$ 进行所有方法的人类研究，并用于生成解释，除非该方法需要特定的数据集来生成解释，例如 Broden [2]。

对于所有使用 SigLIP 指导的方法，我们使用 SigLIP-SO400M-14-386 模型。

B.1.1 基准实现细节

出于实际目的，我们对一些基线方法进行了些许修改。具体的改动细节如下所述：

DnD [1]：最初的实现通过 OpenAI API 使用 GPT-3.5 Turbo。鉴于使用 API 的高成本以及最近的开源 LLM 性能与闭源的 LLM（如 GPT-3.5）相比都很强，我们用 Llama 3.1-8B-Instruct [8] 取代了 GPT-3.5 Turbo。DnD [1] 表明较老的 Llama 2 模型已经比 GPT-3.5 Turbo 更好，并且在神经元描述上与 GPT-4 Turbo 不相上下，所以我们选择 Llama 3.1 不会降低 DnD 相对于 GPT-3.5 Turbo 的性能。

MAIA [23]：相比原始的实现，我们用更新的 GPT4o-2024-08-06 [20] 替换了 GPT4-vision-preview，因为它具有较低的 API 成本和更好的性能。这种方法仍然相当昂贵，为我们生成 ResNet-50 Layer4 和 ViT-B-16 Layer11 MLP 的 100 个随机选定神经元的描述，成本分别约为 $ 65 和 $ 116。请注意，这个费用还包括重复试验，比如对 ResNet 和 ViT 各自 ~ 10 和 ~ 20 个神经元进行实验，这些神经元在第一次运行中没有产生任何描述。起初，我们还尝试了支持视觉输入的开源 LLM（即 VLM），如 Llama-3.2-11B-Vision-Instruct [8]、Llava-OneVision-Qwen2-7B-ov-hf [15] 和 CogVLM2-Llama3-Chat-19B [12]。然而，这些 VLM 与 MAIA 的配合效果不好，因为它们无法生成可执行代码，给它们提供较长提示时，会遗忘图像标记而只专注于最后几个文本标记，并且一次只允许一个图像输入。很可能是因为这些 VLM 倾向于视觉问答，而不具备 GPT4/4o 的更通用功能。

依赖于二维激活的方法：许多方法旨在通过二维激活来解释 CNN 的整个通道 [2, 16, 11, 14]。对于 ResNet-50 的 layer4，我们将预平均池化激活输入这些方法，以便提供适当的二维输入。然而，对于 ViT-B-16 的最后一层，只有 CLS-token 的激活会影响输出，因此我们是在解释具有标量激活的神经元。在这种情况下，我们将标量激活广播到一个所有空间位置值均相同的二维张量中。然而，这并不是使用这些方法的预期方式，这可能部分解释了某些方法在 ViT 神经元上表现较差的原因，正如表 1 和 2 所示。

CCE [14]：对于集群组成解释，我们测试了两个不同版本：$l = 15$ 版本对应于默认版本，对每个 5 个激活集群给出长度为 3 的解释。对于 $l = 5$ 版本，我们使用解释长度 =1，伴随 5 个激活集群。我们还使用 [14] 的实现，通过设置解释长度 =3 和集群数 =1 来重现组成解释 [16] 的结果。

B.1.2 神经元子集

对于表 1 和 2 中的大多数方法，我们报告了 ResNet-50 的 layer4 中所有 2048 个神经元和 ViT-B-16 的 layer11 mlp 中所有 3072 个神经元的平均相关性得分。然而，由于某些方法具有较高的计算和/或 API 成本，我们只能解释这些神经元的一个子集，并在表 1 和 2 中报告这些子集的平均得分。我们报告了以下方法的部分神经元结果：

- MAIA [23]：对于 RN50 和 ViT-B-16，各随机选择 100 个神经元的子集。
- CCE [14] $l = 5$：对于 ViT-B-16，我们使用了 1420 个神经元的子集。对 RN50 进行评估时使用了所有神经元。
- CCE [14] $l = 15$：RN50：984 个神经元的子集。ViT-B-16：422 个神经元的子集。

。其他所有方法均在每一层的所有神经元上进行了评估。

B.2 自动化评估细节

对于我们的自动评估（见第 5.1 节），我们使用 [18] 描述的相关评分模拟。这个评估最初是由 [4] 针对语言模型神经元提出的。

模拟评估的基本思想是使用解释来预测在未见输入上的神经元激活。通过相关评分，我们评估预测激活 s 和实际神经元激活 a_k 在 10,000 个输入的整个测试集上的相关系数 ρ，正如 [18] 所做的那样。

为了简单的解释，预测的激活 s 就是该输入上概念的存在。

$$s(x_i, t) = [c_t]_i \tag{13}$$

对于线性解释，$E = \{(w_1, t_1), ..., (w_l, t_l)\}$ 预测激活 s 按照 [18] 计算为：

$$s(x_i, E) = \sum_{w_j, t_j \in E} w_j [c_{t_j}]_i \tag{14}$$

对于组合解释 [16]，我们使用概率逻辑计算预测激活。不同比较基本的逻辑运算符计算如下：

$$s(x_i, t_1 \text{ AND } t_2) = [c_{t_1}]_i \cdot [c_{t_2}]_i \tag{15}$$

$$s(x_i, t_1 \text{ OR } t_2) = 1 - (1 - [c_{t_1}]_i) \cdot (1 - [c_{t_2}]_i) \tag{16}$$

$$s(x_i, \text{NOT } t) = 1 - [c_t]_i \tag{17}$$

然后通过迭代应用这些规则来计算更复杂成分的预测。

聚类组合解释: CCE [14] 解释的形式为 $E = \{(l_1, u_1, F_1), ..., (l_r, u_r, F_r)\}$，其中 r 是激活簇的数量，l_j, u_j 是该簇的激活的下限和上限，而 F_j 是该簇激活的组合解释。为了基于该解释预测神经元激活，我们使用以下公式：

$$s(x_i, E) = \sum_{l_j, u_j, F_j \in E} \frac{l_j + u_j}{2} s(x_i, F_j) \tag{18}$$

这意味着如果根据聚类公式的概念存在，我们预测神经元的激活将位于聚类激活范围的中间。

对于所有自动评估，我们使用 SigLIP-SO400M-14-386 模型预测 c_t，并遵循 [18]。

B.3 定理 1

假设我们正在估计函数 $h(x)$ 在 $x \sim \mathcal{P}$ 时的期望值。令 $\mathcal{X}$ 为 $\mathcal{P}$ 的支集。

Theorem 1 ([22] , 第 3.3.2 节，定理 3.12)**.** *对于采样分布为 q 的重要性采样：*

$$\mathbb{E}_{x \sim \mathcal{P}}[h(x)] = \int_{\mathcal{X}} h(x) \frac{p(x)}{q(x)} q(x) dx \approx \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{h(x_i) p(x_i)}{q(x_i)}.$$

最小化方差的 q 的选择满足 $q(x) \propto |h(x)|p(x)$.

证明：从 [22] 复制而来。

$$Var\left[\frac{h(x)p(x)}{q(x)}\right] = \mathbb{E}_q\left[\left(\frac{h(x)p(x)}{q(x)}\right)^2\right] - \left(\mathbb{E}_q\left[\frac{h(x)p(x)}{q(x)}\right]\right)^2 \tag{19}$$

由于第二项 $\left(\mathbb{E}_q\left[\frac{h(x)p(x)}{q(x)}\right]\right)^2 = \left(\int_{\mathcal{X}} h(x)p(x)dx\right)^2$ 不依赖于 q，为了最小化方差，我们只需要最小化第一项。

根据詹森不等式可得：

给了我们第一项的下界。如果我们设置

$$q(x) = \frac{|h(x)|p(x)}{\int_{\mathcal{X}} |h(z)|p(z)dz} \tag{20}$$

这是一个有效的概率分布，我们得到：

$$\mathbb{E}_q\left[\left(\frac{h(x)p(x)}{q(x)}\right)^2\right] = \left(\int_{\mathcal{X}} |h(z)|p(z)dz\right)^2 \tag{21}$$

这恰好匹配了下界，证明通过设置可以达到最小方差

$$q(x) = \frac{|h(x)|p(x)}{\int_{\mathcal{X}} |h(z)|p(z)dz} \propto |h(x)|p(x) \tag{22}$$

B.4 对均匀先验中 β 的敏感性

我们的评级聚合方法 3a Bayes - 均匀采样，如第 **??** 节所述，依赖于 β 超参数来选择均匀先验。在本节中，我们在设置 2（MTurk 测试）中进行了测试，比较了 β 的不同值，并显示它对相关误差的影响非常小，如图 4 所示。在我们的实验中，我们使用了 $\beta = 0.01$，其整体表现良好，但正如我们所见，具体选择对结果几乎没有影响。我们认为这主要是因为改变先验对预测的 c_t 有相对线性的影响，而且由于相关系数标准化了 c_t，c_t 的尺度不会改变相关性。

B.5 计算细节

我们的主要贡献集中在高效的众包评估，因此我们的方法在计算方面并不昂贵。与我们的方法相关的主要计算成本是为整个 $\mathcal{D}$ 计算 SigLIP 图像编码器的输出，因为这些输出对于重要性采样和贝叶斯-SigLIP 先验都是必需的。然而，这只是一次相对便宜的成本，大约需 20 分钟单一 NVIDIA RTX 6000 Ada Generation GPU 即可完成。

本文的主要计算开销涉及运行基线方法。在下面的表格 4 和 5 中，我们报告了使用单个 NVIDIA RTX 6000 Ada Generation GPU，并使用不同描述方法来解释一层的所有神经元的近似运行时间。

Simple Exp. (l =1)	ND [3]	MILAN [11]	CD [17]	INVERT l=1 [5]	DnD [1]	MAIA [23]	LE(label) l=1 [18]	LE(SigLIP) l=1 [18]
RN-50 (Layer4)	$\sim$ 1 hr	$\sim$ 1 hr	$\sim$ 5 mins	$\sim$ 1 hr	$\sim$ 55 hrs	$\sim$ 255 hrs	$\sim$ 1 hr	$\sim$ 1 hr

Table 4: 不同 $l = 1$ 基线方法解释神经元的近似运行时间。

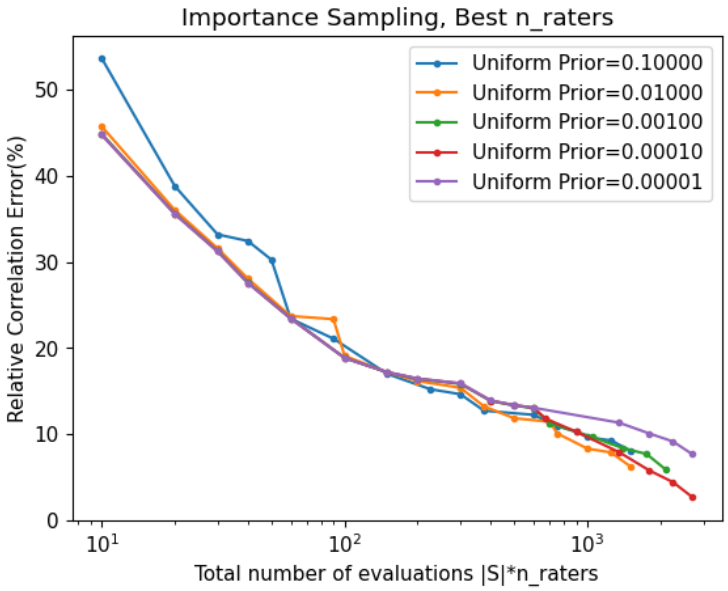


Figure 4: 比较使用不同 β 值的均匀先验的错误率。

Complex Exp. (l >1)	Comp Exp [16] l=3	INVERT [5] l=3	CCE [14], l= 5	CCE [14], l=15	LE(label) [18] l=4.37/1.97	LE(SigLIP) [18] l=4.66/1.82
RN-50 (Layer4)	$\sim$ 92 hrs	$\sim$ 24 hrs	$\sim$ 87 hrs	$\sim$ 275 hrs	$\sim$ 1 hr	$\sim$ 1 hr

Table 5: 不同复杂解释基线方法的大致运行时间。

C 附加图表

图 5 和 6 展示了示例神经元以及由不同解释方法分配的描述，以及我们众包研究对这些解释估计的相关性得分。我们根据估计的相关系数对解释进行着色，其中绿色代表 $\rho \geq 0.25$ ，黄色代表 $0.25 > \rho \geq 0.15$ ，红色代表 $0.15 > \rho$ 。

MTurk 实验详情：图 7 展示了显示给评价者的完整用户界面。我们选择了居住在美国的评价者，他们的任务通过数量超过 10,000 个且任务通过率为 $> 98\%$ 。每位评价者每个任务获得 $ 0.05 的报酬（我们还需为每个任务支付 MTurk $ 0.01 的费用），根据我们的测试，每个任务需大约 15 秒完成，估算时薪为 $ 12。

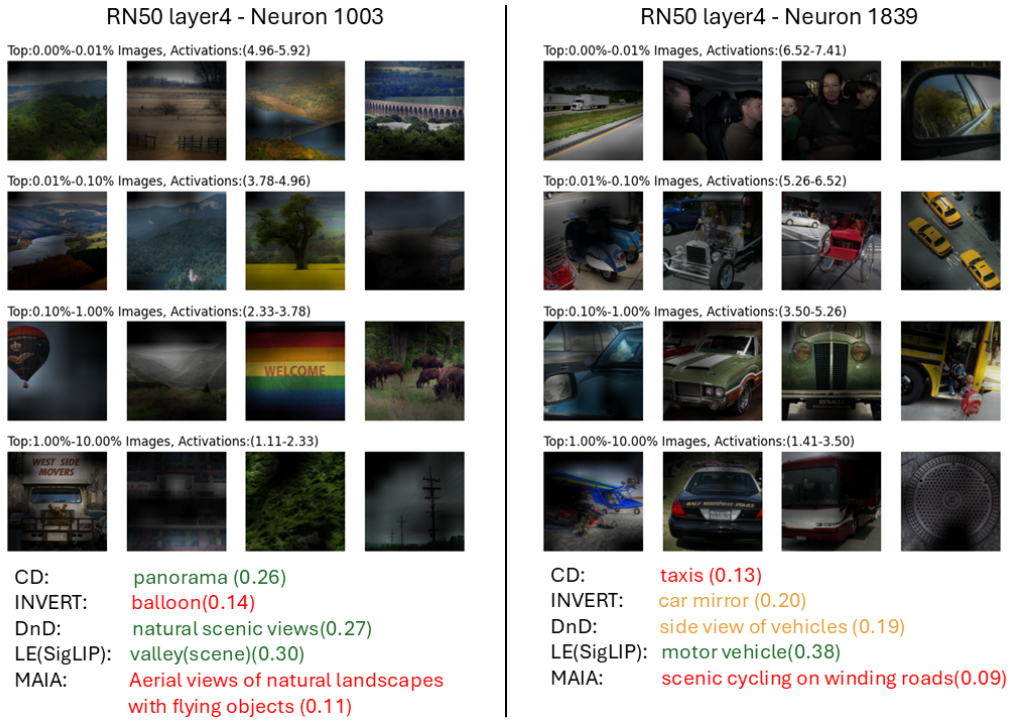


Figure 5: 示例神经元的可视化、其描述以及来自我们众包评估中的相关性分数（使用贝叶斯和 SigLIP 先验）。我们根据相关性分数对描述进行了着色。

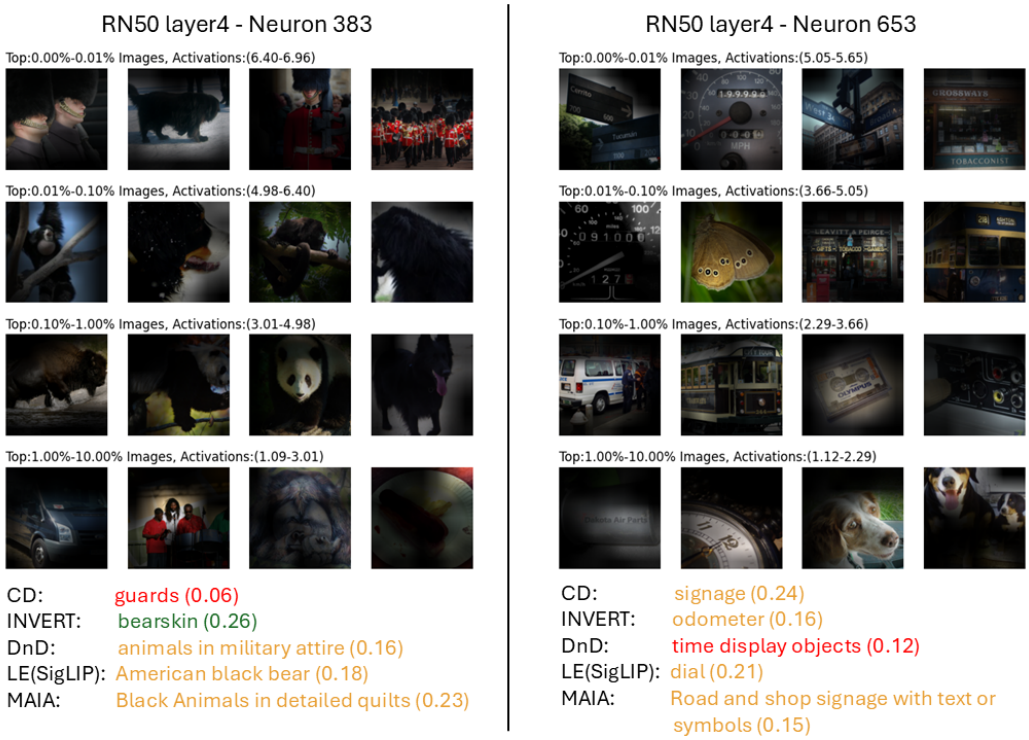


Figure 6: 来自我们众包评估（使用 SigLIP 先验的贝叶斯方法）的示例神经元、其描述和相关性评分的可视化。我们根据相关性评分为描述进行了着色。

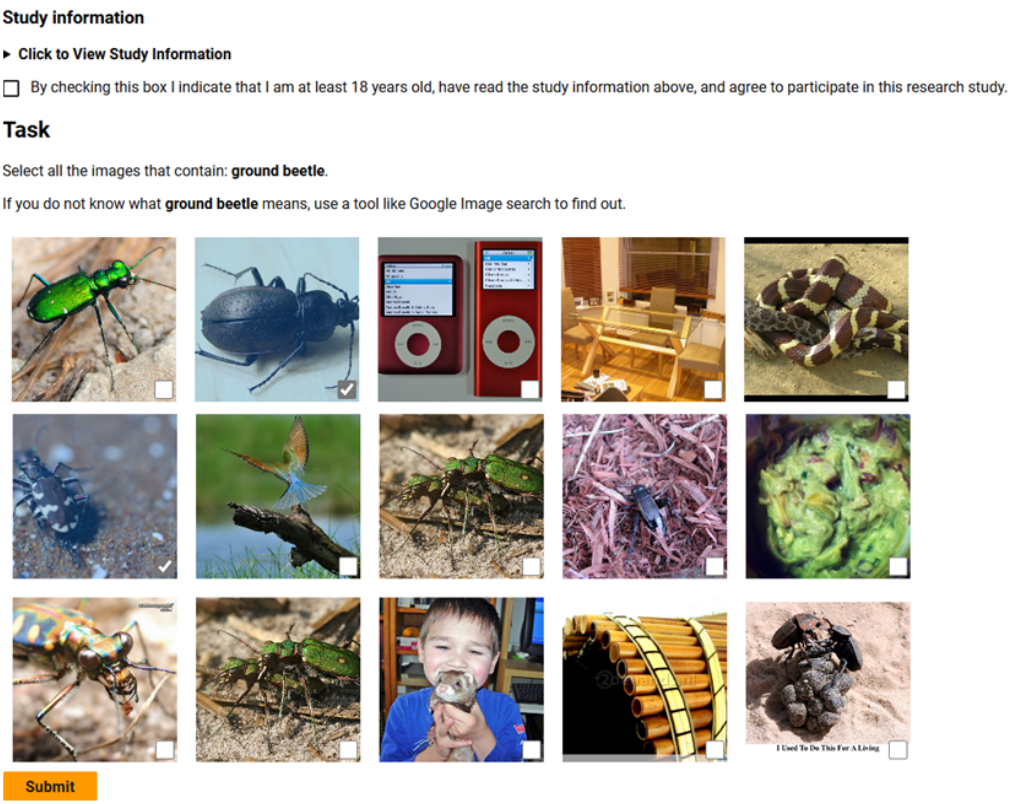


Figure 7: 我们的用户研究界面。