关于修改神经网络的感知

Manuel de Sousa Ribeiro，**Jo˜ao Leite**NOVA LINCS，
里斯本新大学，葡萄牙
mad.ribeiro@campus.fct.unl.pt，jleite@fct.unl.pt

arXiv:2303.02655v1 [cs.AI] 5 Mar 2023

摘要

人工神经网络已被证明是极其有用的模型，推动了人工智能及许多其他领域近期的多项突破。然而，由于人类难以理解这些模型如何得出结果，它们通常被视为黑箱。在本研究中，我们提出一种方法，可修改人工神经网络对特定人类定义概念的感知，从而生成假设情景，帮助人们理解甚至调试神经网络模型。我们通过实证评估，在合成数据集和 ImageNet 数据集上对不同模型测试了所提出的方法，评估模型是否能正确理解所实施的操控，并分析它们对此的反应。

1 引言

本文探讨如何改变神经网络对特定人类定义概念的感知，这些概念并未直接编码在其输入中，目的是更好地理解此类概念如何影响模型的预测结果。我们的研究结果表明，只需少量标注数据即可实现这一点，而且无需更改或重新训练现有的神经网络模型。

在过去几十年中，人工神经网络推动了多个领域的重大进展，涵盖文本、图像、视频和语音处理 [Khan 等、 2020], 医学[Shahid 等、 2019], 经济学和金融 [Ghoddusi 等以及许多其他领域。近年来，人们已实现众多基于神经网络的成功应用 [Abiodun等，使这些模型无处不在。

尽管神经网络广受欢迎，但它们缺乏可解释性，因为它们无法直接提供人类可理解的指示，说明为何会得出某个结果 [Hitzler 等人， 2020]。这促使人们开发了多种方法来解决这一缺陷（参见第 6 节）。最常见的方法通常侧重于指出哪些输入对输出的贡献最大，或用一个本质上可解释的模型替代原模型 [Kim 等人， 2018]。然而，这些方法把理解所提供的解释为何能支持该输出的负担留给了用户，例如，用户必须理解为何某一组特征（如图像中的像素）及其取值会导出该输出。多项用户研究 [Adebayo 等人，2020；Chu 等人，2020；Shen 和 Huang， 2020] 表明，这些方法给出的解释往往会被最终用户忽视，或对他们毫无帮助。

原因之一是，人类通常不会以非常低层次的特征进行推理，例如图像中的单个像素；人类通常依据更高层次的人类定义概念进行推理。例如，从人的角度出发，要解释网络为何将某张火车图片分类为客运列车，一种有用的解释可能会指出车厢上有窗户，而不是指出图像中的特定像素。

此外，当人类试图确定某些复杂现象的原因时，往往会借助反事实推理 [Miller，2019], ，尝试理解不同场景如何导致不同结果。这种方法似乎也有助于解读人工神经网络，因为它强调了原因——发生了什么变化——以及结果——变化导致了什么改变。例如，为了更好地理解神经网络如何对一张特定的火车图片进行分类，用户可以询问：如果图片中包含一节客运车厢，模型会输出什么。我们希望能够根据不同人类定义的概念对模型的预测所产生的影响，生成此类反事实场景。使用人类定义的概念（前例中的客运车厢概念）十分重要，因为这些概念为描述反事实提供了语义，同时也确保它们对特定模型的用户而言有意义且易于理解。

已有研究致力于开发为人工神经网络生成反事实的方法 [Guidotti,2022], 其中一些方法甚至能够针对特定属性生成反事实样本 [Yang 等，2021]。然而，据我们所知，这些方法关注的是输入样本的特定变化可能如何影响模型的输出，却忽略了模型实际上从这些样本中感知到了什么。此外，当前生成反事实的方法通常实现起来较为复杂，往往需要训练额外的模型来生成反事实 [Stepin 等，2021], 并使用特定的神经网络架构，例如可逆神经网

络 [Hvilshøj 等, 2021]。

在这项工作中，我们从不同的抽象层次探讨反事实生成问题，重点关注模型从给定输入中感知到了哪些特定的人类定义概念，以及这对模型输出有何影响。我们的想法是“说服”模型，使其认为自己感知到了某个特定概念——例如客运车厢——而不生成包含该概念的具体图像，并观察这对模型输出的影响。我们不再专注于生成特定的反事实样本，而是转而生成关于模型如何感知人类定义概念的反事实，从而能够以人类可理解的方式，更好地了解神经网络中编码的信息如何影响模型的预测。通过操控模型对特定感兴趣概念的感知，用户可以探索神经网络的输出如何取决于其对这些概念的感知。

要基于模型感知到的内容，而不是输入给模型的内容来生成反事实，我们需要能够理解神经网络模型感知到了什么，并能够操控模型对特定概念的感知。我们的灵感来自神经科学领域的研究，其中可以识别出高度选择性、似乎能够表征特定刺激含义的神经元 [Reddy and Thorpe, 2014]。这些神经元被称为概念细胞，似乎“提供了对概念稀疏、显式且不变的表征” [Quiroga 等, 2005; Quiroga, 2012]。这些细胞的发现有助于更好地理解大脑——一个复杂的、亚符号系统——如何运作，以及它如何关联不同概念[Gastaldi 等, 2022]。此外，通过光遗传学技术，我们可以操控特定神经元的活动，并了解其功能 [Okuyama, 2018]。

类似地，如果能够为神经网络中的特定神经元赋予含义，我们就能更好地理解给定模型中编码了哪些信息，以及它可能如何关联不同概念。此外，鉴于通常可以访问神经网络的内部结构，我们可以操纵那些已被赋予含义的神经元的输出，并考察此类变化在不同情境下如何影响模型，从而生成反事实场景。

基于现有证据——人工神经网络中似乎会出现与概念细胞性质相似的神经元 [Goh等人, 2021], 我们提出这样的假设：通过识别神经网络中哪些神经元充当特定人类定义概念的概念细胞，并操纵这些神经元的激活值，我们应当能够改变神经网络对这些概念的感知。

本文提出并测试了一种方法，通过改变人工神经网络对特定人类定义的概念的感知，生成反事实情景。在第 2 节中，我们介绍所提出的方法：第 2.1 节讨论如何确定神经网络中哪些神经元能够识别特定概念；第 2.2 至 2.4 节则探讨所提方法的不同方面，并提供实验证据以支持我们的主张。第 3 节和第 4 节展示了所提方法的潜在应用。第 5 节中，我们在 ImageNet 数据集上应用并测试该方法。第 6 节讨论相关研究，并在第 7 节作出总结。

2 操纵神经网络感知的方法

为了针对神经网络模型对人类定义概念的感知生成反事实，我们提出一种由三个主要步骤组成的方法。对于每个关注的概念：

a) 估计每个神经元对该概念的敏感程度，即其激活值区分该概念存在的样本与该概念不存在的样本的效果有多好；b) 根据神经元敏感度值，选出被视为“类似概念细胞的”神经元，我们将其称为概念神经元；c) 对于每个概念神经元，计算两个激活值，分别表示该神经元在该概念存在和不存在的样本上的输出。

考虑一个神经网络模型 $\mathcal{M}: \mathbb{I} \to \mathbb{O}$，它是正在分析的模型，其中 $\mathbb{I}$ 是输入数据空间，$\mathbb{O}$表示模型输出空间。对于每个纳入考虑、由人定义的概念 C，我们假设有一组正 $P_C \subseteq \mathbb{I}$ 样本和一组负 $N_C \subseteq \mathbb{I}$ 样本，分别表示该概念存在和不存在。设 $a_i^{\mathcal{M}}: \mathbb{I} \to \mathbb{R}$ 表示按照某种任意的神经元排序，神经网络模型 $\mathcal{M}$ 中第i^{th} 个神经元的输出。

我们方法的第一步是估计每个神经元对每个纳入考虑的概念的敏感程度，即每个神经元的激活值区分概念存在的样本与概念不存在的样本的效果有多好。我们用 $r_i^{\mathcal{M}}: 2^{\mathbb{I}} \times 2^{\mathbb{I}} \to [0,1]$ 表示函数；对于模型 $\mathcal{M}$中的给定神经元i，该函数接收集合 P_C 和 N_C，并给出一个值，用以表示 i 对概念 C 的敏感程度。在第 2.1 节中，我们将介绍该函数的不同实现方式。

第二步是针对每个感兴趣的概念 C，选择 $\mathcal{M}$ 中哪些神经元应被视为概念 C 的概念神经元。令 $s_C: \mathbb{R} \to \{0,1\}$ 为阈值函数，用于指示：对于给定的概念 C 和敏感度值，该值是否足够高，可以将其视为概念细胞。随后，对于每个感兴趣的概念 C，我们将选定神经元的集合确定为 $S_C = \{i: s_C(r_i^{\mathcal{M}}(P_C, N_C)) = 1, i \in [1,k]\}$，其中 k 是 $\mathcal{M}$ 中神经元的数量。在所有实验中，我们通过逐步降低阈值来设定阈值，同时针对每个概念，在包含 100 个样本的验证集上测试模型性能。当通过降低阈值添加了 3 个以上的神经元，且验证集上的性能没有提升时，我们便将阈值设为找到的最佳值。

最后，对于每个被选为某个概念 C 的概念神经元的神经元，我们会计算两个激活值，分别表示该概念存在和不存在时的情况。我们考虑一个函数 $c_i^{\mathcal{M}}: 2^{\mathbb{I}} \to \mathbb{R}$，用于计算 $\mathcal{M}$ 中 i^{th} 神经元的激活值，并用它计算一个表示概念 C 存在的激活值 $c_i^{\mathcal{M}}(P_C)$，以及一个表示该概念不存在的激活值 $c_i^{\mathcal{M}}(N_C)$。在

TypeA $\equiv$ WarTrain $\sqcup$ 空列车 $\exists$has.货运车厢 $\sqcap$ $\exists$has.客车 $\sqcap$ $\exists$has.空车厢 $\sqsubseteq$ 混合列车 TypeB $\equiv$ 客运列车 $\sqcup$ 长货运列车 $\exists$has.(客车 $\sqcap$ 长货车) $\sqcup$ ($\geq$ 2 has. 客车) $\sqsubseteq$ 客运列车 TypeC $\equiv$ 乡村列车 $\sqcup$ 混合列车 $\exists$has. 加固车厢 $\sqcap$ $\exists$has. 客车 $\sqsubseteq$ WarTrain 长货运列车 $\equiv$ 长列车 $\sqcap$ FreightTrain ($\geq$ 2 has.长货车) $\sqcup$ ($\geq$ 3 has.车厢) $\sqsubseteq$ 长列车 空列车 $\equiv$ $\forall$has.(空车厢 $\sqcup$ 机车) $\sqcap$ $\exists$has.空车厢 ($\geq$ 2 has.货运车厢) $\sqsubseteq$ FreightTrain

图1：XTRAINS 数据集本体中的一部分公理，描述列车的表征如何分类。



图2：XTRAINS 数据集的样本图像。

第2.2节中，我们讨论 $c_i^{\mathcal{M}}$ 的不同实现，并比较它们对该方法结果的影响。

随后，当某个输入 $x \in \mathbb{I}$ 被输入神经网络 $\mathcal{M}$ 时，为了生成概念 C 存在或不存在的反事实场景，我们只需将 S_{C} 中每个神经元 i 的激活值 $a_i^{\mathcal{M}}(x)$ 分别替换为 $c_i^{\mathcal{M}}(P_{\mathsf{C}})$ 或 $c_i^{\mathcal{M}}(N_{\mathsf{C}})$。我们将这一步称为向神经网络模型中注入一个概念。

为了测试我们的方法，我们采用可解释抽象列车数据集（XTRAINS） [de Sousa Ribeiro 等人， 2020], 这是一个合成数据集，由列车表征组成，例如图2所示的列车。该数据集包含关于各种视觉概念的标签，以及描述这些概念之间关系的基于逻辑的本体。图1展示了该数据集配套本体的一部分，说明不同概念之间的关系。例如，该本体规定 TypeA 要么是 WarTrain，要么是空列车；而 WarTrain 则包含具有加固车厢和客车的列车。这些概念具有视觉表征，例如，加固车厢呈现为每侧车壁上有两条线的车厢，如图2最左侧图像中的前两节车厢。

在下文所述的实验中，我们采用了 [Ferreira 等人开发的神经网络，该网络经过训练，可识别 TypeA 列车，称为 $\mathcal{M}_{\mathsf{A}}$。该神经网络经过训练，在包含 10 000 张图像的平衡测试集上达到约 99% 的准确率。我们还采用了 [de Sousa Ribeiro 和 Leite 2021] 给出的相关性定义，以确定哪些概念与给定神经网络的任务相关（相对于数据集附带的本体）。

2.1 识别类似概念细胞的神经元

为了能够修改神经网络对特定人类定义概念的感知，首先需要确定模型中哪些神经元正在识别这些概念。根据已有证据，神经网络的内部似乎编码了与其任务相关的概念信息 [de Sousa Ribeiro 和 Leite， 2021], 并且人工神经网络中似乎会出现具有“类似概念细胞”特性的神经元 [Goh 等， 2021], 我们假设神经网络中存在充当概念细胞的神经元，用于处理与其所执行任务相关的概念。本节将探讨如何确定

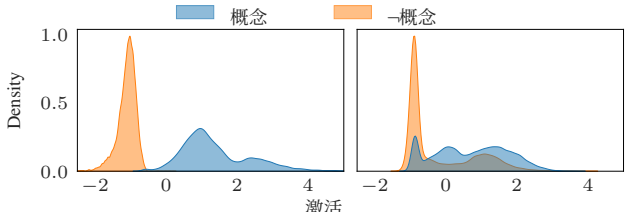

图 3：在给定概念存在或不存在的样本中，两个神经元的概率密度函数。

神经网络模型中的这类神经元。

如果能够根据某个神经元的激活值，将概念 C 存在的样本与概念 C 不存在的样本区分开来，我们就认为该神经元对于概念 C 而言具有“类概念细胞”的特性。图 3 展示了两个神经元在某些 P_{C} 和 N_{C} 集合上的估计概率密度函数。左侧的神经元似乎充当了概念 C 的概念细胞，两个样本集合的分布区分明显。另一方面，右侧的神经元并未充当该概念的概念细胞，因为它的激活值无法区分这两个样本集合。

虽然有一些方法，例如 TCAV [Kim 等人， 2018]以及映射网络 [de Sousa Ribeiro 和 Leite， 2021],能够帮助理解给定模型是否对某个概念敏感，但我们在这里关注的是理解单个神经元是否对该概念敏感。我们的目标不是检验模型是否对某个概念敏感，因为只需部分神经元便可能足以实现这一点；我们要做的是找出所有对该概念敏感的神经元，以便操控它们。

我们考虑 $r_i^{\mathcal{M}}$ 的三种不同实现方式，以评估神经元 i 在 $\mathcal{M}$ 中作为概念细胞的适用性：

- 神经元激活值与数据集标签之间的斯皮尔曼等级相关，计算方式为 $|1 - \frac{6\sum d_j^2}{n\ n^2 - 1()}|$，其中 $d_j = a_i^{\mathcal{M}}(P_{\mathsf{C}j}) - \overline{a_i^{\mathcal{M}}(N_{\mathsf{C}j})}$ 且 $n = |P_{\mathsf{C}}|$，假设 $|P_{\mathsf{C}}| = |N_{\mathsf{C}}|$；
- 线性二元分类器（b）根据神经元的激活值预测数据集标签的准确率，计算方式为 $\frac{|\{x:\ x \in N_{\mathsf{C}} \cap b(a_i^{\mathcal{M}}\ x()) = 0)\}| + |\{x\ x \in P\ \cap b\ a_i^{\mathcal{M}}\ x\ = 1\ \}|}{|N_{\mathsf{C}}| + |P_{\mathsf{C}}|}$: c(()))cc；
- 神经元对 P_{C} 和 N_{C}的激活值的概率密度函数交集，计算方式为 $1 - \int min(f_i^{P_{\mathsf{C}}}(x)\ f_i^{N_{\mathsf{C}}}(x))\ dx$ 其中 $f_i^{P_{\mathsf{C}}}$是神经元 i 对 P_{C} 中正样本的激活值的估计概率密度函数，$f_i^{N_{\mathsf{C}}}$则是对负样本的估计概率密度函数；该方法假设样本独立同分布。

为了检验我们的假设——神经网络通常应该

	斯皮尔曼	准确率	交集
空列车	13	20	16
∃有。客车	15	19	18
∃有。加固车厢	18	22	21
WarTrain	13	15	9

图4：每个概念对应的概念神经元数量 $\mathcal{M}_{\mathsf{A}}$ 。

	斯皮尔曼	准确率	交集
空列车	98.9%	99.1%	**99.6%**
∃有。客车	90.7%	**93.0%**	92.9%
∃有。加固车厢	97.9%	**98.3%**	98.2%
WarTrain	66.6%	65.7%	**99.0%**

图 5：按敏感度指标划分的正确分类样本。

具有与其任务相关的概念对应的概念神经元，我们针对四个随机选取的相关概念计算了上述三项指标：EmptyTrain、∃has。PassengerCar、∃has。ReinforcedCar 和 WarTrain；并针对四个不相关概念进行计算： ∃has。LongWagon、 ∃has。OpenRoofCar、LongFreightTrain 和 MixedTrain。根据我们的假设，对于相关概念，我们应该能够找到在所述指标上取值较高的神经元；而对于不相关概念，则不应该能够找到这样的神经元。对于每项指标，我们都设定了一个阈值，超过该阈值的神经元在本实验中即被视为概念神经元：斯皮尔曼等级相关的阈值为 0。85，线性分类器准确率的阈值为 0。95，概率密度函数交集的阈值为 0。9。

图4展示了针对所有相关概念，按各项指标识别出的概念神经元数量。对于不相关的概念，没有任何指标识别出概念神经元。例如，对于“空列车”这一概念，我们根据斯皮尔曼等级相关识别出13个概念神经元，根据线性分类器的准确率识别出20个。在第2.2节中，我们将探讨所选概念神经元对于我们方法整体性能的重要性。

关于所选概念神经元的数量，虽然具体数量取决于阈值，但值得注意的是， $\mathcal{M}_{\mathsf{A}}$ 包含约 2×10^6 个神经元。因此，正如预期，所选概念神经元的数量与模型中的神经元总数相比微不足道（约为0.001%）。有趣的是，我们发现所有选出的神经元都位于模型的稠密部分；而越抽象、越复杂的概念，越集中在模型较后的稠密层中。

这些结果似乎证实了我们的假设，表明只要人类定义的概念与模型所执行的任务相关，就有可能识别出神经网络模型中激活值与这些概念高度相关的神经元。

2.2 操纵神经网络的感知

现在，我们继续检验主要假设：通过操纵特定概念对应的概念细胞的激活值，可以操纵神经网络对特定人类定义的概念的感知。

我们通过考察这样一个问题来检验这一假设：对于给定样本，如果某个概念被识别，模型的输出理应发生变化；当我们注入该概念时，输出是否确实会发生变化。例如，如果我们向经过 TypeA列车识别训练的 $\mathcal{M}_{\mathsf{A}}$ 输入一张图像，图中有一节客车车厢但没有加固车厢，那么我们预期模型会输出“这不是一列 TypeA 列车”。然而，如果我们识别出 ∃has.ReinforcedCar 的概念神经元——即具有加固车厢这一概念——并修改其输出以表示存在加固车厢，也就是说，我们注入该概念 ∃has.ReinforcedCar，我们预期模型会改变其输出，将其识别为一列 TypeA列车。

为在 XTRAINS 数据集的设定下检验我们的假设，我们依据以下标准，针对第 2.1 节所考虑的四个相关概念中的每一个，将数据集子采样为四个不同的、各含 1000 个样本的集合：

- $S1$ - 包含 ¬TypeA 样本，其中该概念的存在应改变 $\mathcal{M}_{\mathsf{A}}$ 的输出。
- $S2$ - 包含 ¬TypeA 样本，其中该概念的存在不应改变 $\mathcal{M}_{\mathsf{A}}$ 的输出。
- $S3$ - 包含 TypeA 样本，其中该概念的缺失应改变 $\mathcal{M}_{\mathsf{A}}$ 的输出。
- $S4$ - 包含 TypeA 样本，其中该概念的缺失不应改变 $\mathcal{M}_{\mathsf{A}}$ 的输出。

为了确定 $\mathcal{M}_{\mathsf{A}}$ 的输出是否应发生变化，我们依据数据集提供的本体进行推理。

由于我们方法的第一步是计算神经元敏感度，因此，当将我们的方法应用于所描述的四个集合中的每一个时，我们比较第 2.1 节中所述三种指标得到的结果。图 5 展示了在计算概念神经元的激活值时， $\mathcal{M}_{\mathsf{A}}$的输出表现符合预期的样本平均数量；对于 $c_i^{\mathcal{M}}$，将其设为神经元激活值的中位数，同时对于 P_{C} 和 N_{C}，使用 $|P_{\mathsf{C}}| = |N_{\mathsf{C}}| = 1000$。这些结果似乎表明，通过修改特定神经元的激活值，可以操纵神经网络对不同概念的感知，而这些神经元似乎能够识别相应的概念。输出发生变化且符合预期的样本比例较高，也证明了这一点。此外，这些结果还表明，基于神经元针对正样本和负样本的概率密度函数交集的敏感度指标，平均而言似乎能够提供更加一致且更好的结果。因此，之后的所有实验都使用该指标来计算神经元相对于某一概念的敏感度（s_{C}）。

关于计算神经元激活值的方法，$c_i^{\mathcal{M}}$除了像之前的实验一样使用神经元激活值的中位数外，我们还考虑了两种替代方法：计算神经元激活值的众数，以及通过[Tucker *et al.*， 2021]中所述的方法计算这些值。图 6 展示了每种方法所得的平均结果。在所考虑的三项指标中，对于所有概念，使用中位数都取得了最佳结果，正确分类样本数量最多。我们认为，[Tucker *et al.*， 2021] 中的方法之所以结果较差，是因为每个概念神经元所包含的信息具有相当大的冗余性，因此该方法中使用的探针只需借助这些神经元中的一个很小子集就能学会完成任务，这可能导致大多数神经元的激活值保持不变

	中位数	Mode	[Tucker 等人, 2021]
空列车	**99.6%**	99.5%	72.4%
∃有。客车	**92.9%**	92.8%	86.5%
∃有。加固车厢	**98.2%**	97.7%	73.9%
WarTrain	**99.0%**	98.6%	65.3%

图 6：按计算神经元激活的方法划分的正确分类样本。

	$S1$	$S2$	$S3$	$S4$
空列车	99.7%	-	99.5%	99.5%
∃有。客车	99.9%	98.6%	98.7%	74.5%
∃有。加固车厢	99.2%	98.5%	96.2%	99.1%
WarTrain	98.4%	-	100.0%	98.7%

图7：每个集合中正确分类样本的百分比。

在其余所有实验中，我们将 $c_i^{\mathcal{M}}$计算为一个神经元的激活值中位数。

图7展示了在上述各个集合中，注入概念后得到预期 $\mathcal{M}_{\mathsf{A}}$ 输出的样本百分比。总体而言，结果似乎相当积极，表明模型通常会根据所注入的概念输出预期结果。在集合 $S4$ 中注入¬∃PassengerCar的结果似乎明显逊于其余结果，我们认为这表明 $\mathcal{M}_{\mathsf{A}}$学习到了一些伪相关。

注入给定样本后导致模型输出发生预期变化的样本比例很高，这构成了有力证据，表明通过操纵负责识别这些概念的神经元的激活值，可以改变神经网络对于特定人类定义概念的感知。

2.3 所选神经元的重要性

所提出方法的前两个步骤旨在确定模型中的哪些神经元对于给定的目标概念充当概念细胞。随后利用这些神经元操纵神经网络对该概念的感知。因此，人们可能会疑惑，所选概念神经元的具体集合对于该方法的整体性能究竟有多重要。本节讨论改变函数 s_C的阈值所产生的结果。

为了了解方法的性能对概念神经元数量的依赖程度，我们针对每个概念测量模型输出与所执行操作一致的样本比例。图8展示了第2.2节中描述的所有集合在不同数量的概念神经元下所得的平均结果。

我们的结果表明，如果函数 s_C 过于严格，排除了许多充当概念细胞的神经元，那么
少数被选中的神经元可能无法影响模型，也无法持续在模型中产生预期结果——这一点可以从最初增加概念神经元数量时性能显著提升看出。然而，如果函数 s_C 过于宽松，开始纳入那些对于给定概念而言并非有效概念细胞的神经元，性能便会开始下降——如纳入的概念神经元数量较多时性能急剧下降所示。

正如预期，所提出方法的性能很大程度上取决于所选概念神经元集合。这些

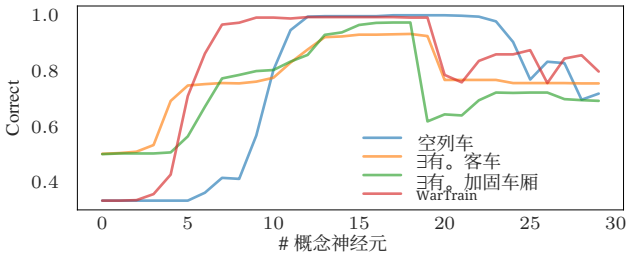

图8：按选定概念神经元数量统计的正确样本数。

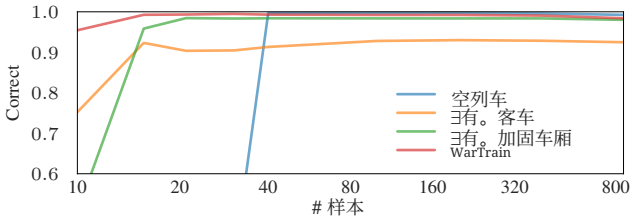

图9：按可用标注数据量统计的正确样本数。

结果表明，为每个感兴趣的概念调整函数 s_C 的阈值有多么重要；这可以通过简单的网格搜索和交叉验证来完成。

2.4 反事实成本

迄今为止的结果表明，通过修改那些似乎负责识别特定概念的神经元的激活，可以操控神经网络对人类定义概念的感知。然而，这些结果假设存在充足的标注数据，并使用总计 2000 个样本来计算神经元敏感度值，以及针对每个概念需要注入的激活值。本节将评估在考虑所需数据量的情况下，所提出的方法是否实用。

为了评估应用该方法的成本，以及可用标注数据量对我们正确修改网络感知能力的影响，我们针对每个注入概念，在使用不同数量的可用数据时，比较第2.2节定义的4个样本集合中获得的平均结果。在整个比较过程中，我们保持用于验证 s_C 阈值的样本数量不变，使用100个样本进行验证。

图9展示了针对每个考察的概念，在注入给定概念后正确分类样本的平均比例。可以看出，即使只有40个标注样本——其中20个样本包含该概念，另外20个不包含——也能相当成功地操控神经网络对每个考察概念的感知。

当标注数据量不足以正确识别给定概念对应的概念神经元时，该方法的性能会开始显著下降。这似乎表明，只要有足够的标注数据来识别模型中哪些神经元对我们想要注入的概念敏感，我们就应该能够操控神经网络对该概念的感知。

3 解读神经网络

到目前为止，我们描述了一种通过修改充当人类定义的概念的概念细胞的神经元激活值，为神经网络模型生成反事实的方法。通过这种方法，我们可以理解不同概念如何影响模型输出。如果我们希望理解模型如何关联那些未在模型输出中体现的不同概念，该怎么办?

例如，人们可能想了解 $\mathcal{M}_A$ 是否已经学会了空列车没有客车，以及如果 $\exists$有.客车，那么这列火车就不是空列车。尽管这两个概念都没有体现在 $\mathcal{M}_A$的输出中，但它们都与其任务相关，因此人们可能会想确认该模型是否能够理解这两个概念之间的关系。

为了了解模型是否识别了某个未体现在其输出中的概念，我们使用映射网络 [de Sousa Ribeiro 和 Leite, 2021]。这些小型神经网络经过训练，可通过神经网络模型的激活值判断其是否识别了某个人类定义的概念。

为了验证 $\mathcal{M}_A$ 是否已经学会了这两个概念之间所描述的关系，我们针对空列车和 $\exists$有.客车这两个概念训练了两个映射网络；二者在由 1000 个样本组成的均衡测试集上的准确率均超过 99%。为了测试 $\mathcal{M}_A$ 是否已经学会空列车没有客车这一点，我们在 1000 个载有客车的列车样本上注入空列车概念。通过针对 $\exists$有.客车训练的映射网络，我们验证了注入后，在 95.6% 的样本中， $\exists$有.客车这一概念从存在变为不存在。这有力地表明，该模型已经学会空列车没有客车。

同样，我们测试 $\mathcal{M}_A$ 是否学会了“如果一列火车有一节客车，那么它就不是空列车”。我们在 1000 个空列车样本上注入概念 $\exists$has.PassengerCar，并观察到其中只有 2% 的样本中，空列车这一概念消失了。这表明 $\mathcal{M}_A$ 尚未学会“如果一列火车有一节客车，那么它就不是空列车”。

这个例子展示了我们的方法的一种可能应用，即调查模型是否编码了用户定义概念之间的某些关系。但更重要的是，模型似乎能够自然地整合注入的信息，并影响相关概念，就仿佛模型真的感知到了注入的概念一样。

4 纠正神经网络的误解

当神经网络模型给出错误结果时，很难判断错误的原因。错误往往可能是因为模型未能正确感知输入所表示内容的一部分，在这种情况下，可以使用所提出的方法来“纠正”模型的错误结果。

为了检验我们的假设，我们选取 XTRAINS 中所有 $\mathcal{M}_A$ 给出错误结果的样本。我们为概念 $\exists$has.ReinforcedCar 和$\exists$has.PassengerCar 训练映射网络——其在一个包含 1000 个样本的平衡测试集上的准确率超过 99%——并用它们来识别



图 10：“罗得西亚脊背犬面部”的图像。

$\mathcal{M}_A$ 给出错误结果的原因是否是模型未能识别本应识别的这两个概念之一。

我们选取所有 $\mathcal{M}_A$ 给出假阴性结果的样本，即输入样本实际上属于 TypeA，却被判定为不属于。从这些样本中我们发现，对于未识别出概念 $\exists$中的。ReinforcedCar、但输入中确实存在该概念的样本，注入 $\exists$中的。ReinforcedCar 后，96。1% 的样本输出得到纠正。同样，对于未识别出 $\exists$中的。PassengerCar、但输入中确实存在该类别的样本，注入该类别后，98。7% 的样本输出得到纠正。

这些结果表明，即使模型给出错误结果，我们的方法仍然适用，使人们能够检验是否是不同的概念导致了该结果。这使用户能够进一步了解模型为何产生错误输出，并找出模型或其训练中可能存在的缺陷。

5 使用真实世界数据进行验证

到目前为止，我们只考虑了一个合成数据集。现在，我们将探讨能否用真实数据复现类似结果。为进行这项验证，我们考虑 ImageNet 数据集 [Russakovsky 等人， 2015]，并考察 [Sandler 等人提出的预训练MobileNetV2模型2018]。

所提出的方法基于以下假设：通过识别模型中哪些神经元负责识别人类定义的某个概念，我们能够通过修改这些神经元的激活值来改变模型对该概念的感知。因此，我们的验证重点在于能否识别负责某个人类定义概念的神经元，以及能否通过注入该概念来改变模型的输出。

为了在这种设置下测试我们的方法，我们选取了一个随机输出类别：“罗得西亚脊背犬”——一种特定的犬种；并从 ImagenetNet 中选出98张图像，以此定义“罗得西亚脊背犬面部”这一概念。这些图像中都能看到罗得西亚脊背犬，且其面部位于画面中央并完整可见，如图10所示的样本所示。我们选择这一概念，是基于这样的假设：它有助于模型对罗得西亚脊背犬图像进行分类。

为了测试注入这一概念是否能够改变模型的感知，我们使用了全部329个样本——这些样本来自训练集和验证集，其中模型在考虑其Top-1结果时，将罗得西亚脊背犬类别错误地预测为阴性。注入“罗得西亚脊背犬面部”这一概念后，其中48%的样本输出了罗得西亚脊背犬类别。如果我们考虑模型在Top-5



图 11：经过审查的“罗得西亚脊背犬面部”图像。

	未审查	已审查	已审查 + 注入
Top-1	70%	36%	64%
Top-5	100%	64%	88%

图 12：模型在“罗得西亚脊背犬”类别上的准确率。

输出中错误地预测为阴性的38个样本，我们成功将模型输出中的71.1%修改为罗得西亚脊背犬类别。

由于“罗得西亚脊背犬面部”这一概念本身不足以使模型输出“罗得西亚脊背犬”类别，因此这些结果证明模型确实在使用这一概念，而且通过操纵其概念神经元，我们能够修改模型输出。

我们进一步检验以下假设：“罗得西亚脊背犬面部”这一概念对于模型识别该犬种至关重要，并且我们能够通过注入该概念生成反事实样本。为此，我们首先按照图 11 所示，遮蔽 ImageNet 验证集中所有“罗得西亚脊背犬面部”。随后，我们测量遮蔽犬只面部前、遮蔽后，以及在遮蔽后的图像中注入“罗得西亚脊背犬面部”概念后的模型准确率。图 12 所示结果表明，遮蔽犬只面部会导致准确率显著下降，这似乎证明了该概念对模型的重要性。之后，即使面部已被遮蔽，我们仍然注入该概念，并观察到准确率大幅上升，这表明注入概念成功补充了遮蔽所移除的部分信息，并帮助模型给出正确分类。

这些结果表明，即使在使用真实数据、中等规模的模型（$\approx 7\times10^6$ 个神经元）且标注数据较少的情况下，也可以识别出模型中哪些神经元充当给定概念的概念细胞；更重要的是，通过操控这些神经元的激活值，可以操控神经网络对该概念的感知。

6 相关工作

过去几年中，解释人工神经网络的方法不断发展，其中代理方法是最受欢迎的方法之一（参见 [Gilpin 等，2018]）。这类方法旨在用一个本身具有可解释性且行为相似的模型来替代待解释的模型。相比之下，我们生成反事实场景的方法无需更改或替换原始模型，而这并非总是可行的。

另一种广受欢迎的方法是显著性与归因方法（参见 [Li 等，2021]），该方法通过为每个输入特征分配一个贡献值来解释模型的行为，这个贡献值表示相应输入特征对给定预测的贡献。尽管这些方法可能有助于了解哪些特征对给定预测的贡献最大，但它们无法阐明为何这些贡献能够证明该输出是合理的，因此，理解这些贡献与具体输出之间的关联这一负担便留给了用户。

一些基于反事实的方法（参见 [Guidotti，2022]）也存在类似问题：它们提供反事实样本，却没有解释为什么该样本会导致不同的结果。我们认为，反事实方法若能跳出仅生成特定反事实样本的做法，转而以人类可理解的方式说明这些反事实如何影响模型并导致某个其他特定输出，将会更有裨益。

近年来，神经符号方法致力于理解模型是否对特定的人类定义的概念 [Kim 等，2018], 敏感，以及如何利用模型中这些概念的表征来解释模型的输出 [de Sousa Ribeiro 和 Leite，2021], ；同时也探讨如何生成人类可理解的理论，以表示模型的分类过程 [Ferreira 等，2022]。本文通过一种聚焦于模型所感知内容的方法，将这两种研究路径联系起来；该方法能够针对人类定义的概念，通过操纵模型所感知的内容来生成反事实。

7 结论

本文提出了一种方法，用于生成关于神经网络模型对人类定义的概念感知内容的反事实，使用户能够询问模型，并了解模型的输出如何取决于这些概念。为此，我们探讨了如何识别模型中哪些神经元对给定概念敏感，以及如何利用这些神经元操纵模型，使其感知该概念。通过实验评估，我们表明，只需少量标注数据，就能操纵神经网络对不同概念的感知，而且无需更改或重新训练原始模型。

我们对所提出的方法进行了形式化描述，并说明了如何应用该方法生成反事实，以及如何利用它更好地理解模型中不同概念之间的关联，并检查和纠正模型的错误分类。

我们得出结论：对于与神经网络任务相关的概念，通常存在少数几个特定神经元对该概念进行编码，并负责在模型中识别该概念。修改这些神经元的激活值，类似于刺激人脑中的概念细胞，可以欺骗模型，使其认为自己正在感知该概念。这种简单的方法似乎能够有效操纵高层抽象概念，让用户无需付出太多努力，就能探索模型在不同情境下会如何响应。

未来，我们希望探索如何利用该方法来搜索习得偏差，并找出概念之间缺失或不理想的关联。

致谢

作者谨此感谢 FCT 通过博士奖学金（UI/BD/151266/2021）及strate-提供的支持。

gic 项目 NOVA LINCS (UIDB/04516/2020)。

参考文献

[Abiodun等， 2018] Oludare Isaac Abiodun、Aman Jan- tan、Abiodun Esther Omolara、Kemi Victoria Dada、Nachaat AbdElatif Mohamed 和 Humaira Arshad。《人工神经网络应用领域的最新进展：综述》。*Heliyon*，4(11):e00938，2018年。

[Adebayo等， 2020] Julius Adebayo、Michael Muelly、Ilaria Liccardi 和 Been Kim。《模型解释调试测试》。载于 Hugo Larochelle、Marc'Aurelio Ran- zato、Raia Hadsell、Maria-Florina Balcan 和 Hsuan-Tien Lin 编，《神经信息处理系统进展 *33*：神经信息处理系统年会 *2020*，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上》，2020年。[Chu等， 2020] Eric Chu、Deb Roy 和 Jacob Andreas。《视觉解释有用吗？模型参与预测的案例研究》。*CoRR*，abs/2007.12248，2020年。[de Sousa Ribeiro 和 Leite， 2021] Manuel de Sousa Ribeiro 和 Jo˜ao Leite。《面向可解释人工智能的人工神经网络与本体对齐》。载于*AAAI'21* 会议论文集，第4932–4940页。AAAI Press，2021年。[de Sousa Ribeiro等， 2020] Manuel de Sousa Ribeiro、Ludwig Krippahl 和 Jo˜ao Leite。《可解释抽象列车数据集》。*CoRR*，abs/2012.12115，2020年。[Ferreira等， 2022] Jo˜ao Ferreira、Manuel de Sousa Ribeiro、Ricardo Gonc.alves 和 Jo˜ao Leite。《窥探黑箱：神经网络的基于逻辑的解释》。载于 Gabriele Kern-Isberner、Gerhard Lakemeyer 和 Thomas Meyer 编，《第*19*届知识表示与推理原理国际会议论文集，*KR 2022*，以色列海法，*2022*年*7*月*31*日至*8*月*5*日》，2022年。[Gastaldi 等， 2022] Chiara Gastaldi、Tilo Schwalger、Emanuela De Falco、Rodrigo Quian Quiroga 和 Wulfram Gerstner。《共享概念细胞支持关联时：重叠记忆印迹理论》。*PLOS* 计算生物学，17(12):1–44，2022年12月。[Ghoddusi 等， 2019] Hamed Ghoddusi、Germ´an G. Creamer 和 Nima Rafizadeh。《机器学习在能源经济学与金融中的应用：综述》。*EnergyEconomics*，81:709–727，2019年。[Gilpin等， 2018] Leilani H. Gilpin、David Bau、Ben Z. Yuan、Ayesha Bajwa、Michael A. Specter 和 Lalana Ka- gal。《解释解释：机器学习可解释性综述》。载于 Francesco Bonchi、Foster J. Provost、Tina Eliassi-Rad、Wei Wang、Ciro Cat- tuto 和 Rayid Ghani 编，《第*5*届 *IEEE* 数据科学与高级分析国际会议，*DSAA 2018*，意大利都灵，*2018*年*10*月*1*日至*3*日》，第80–89页。IEEE，2018年。[Goh 等， 2021] Gabriel Goh、Nick Cammarata、Chelsea Voss、Shan Carter、Michael Petrov、Ludwig Schubert、Alec Radford 和 Chris Olah。《人工神经网络中的多模态神经元》。Distill，2021年。https://distill.pub/2021/multimodal-neurons.[Guidotti， 2022] Riccardo Guidotti。《反事实解释及其寻找方法：文献综述与基准测试》。《数据挖掘与知识发现》，2022年4月。[Hitzler 等人， 2020] Pascal Hitzler、Federico Bianchi、Monireh Ebrahimi 和 Md. Kamruzzaman Sarker。《神经符号集成与语义网》。《语义网》，11(1):3–11，2020年。[Hvilshøj 等人， 2021] Frederik Hvilshøj、Alexandros Iosifidis 和 Ira Assent。《ECINN：基于可逆神经网络的高效反事实》。《第32届英国机器视觉会议 2021》，BMVC 2021，线上，2021年11月22日至25日，第43页。BMVA Press，2021年。[Khan 等人， 2020] Asifullah Khan、Anabia Sohail、Umme Zahoora 和 Aqsa Saeed Qureshi。《深度卷积神经网络近期架构综述》。《人工智能评论》，53(8):5455–5516，2020年。[Kim 等人， 2018] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie J. Cai、James Wexler、Fernanda B. Vi´egas 和 Rory Sayres。《超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试》。载于 Jennifer G. Dy 和 Andreas Krause（编），《第35届国际机器学习大会论文集》，ICML 2018，Stockholmsm¨assan，斯德哥尔摩，瑞典，2018年7月10日至15日，《机器学习研究论文集》第80卷，第2673–2682页。PMLR，2018年。[Li 等人， 2021] Xiao-Hui Li、Yuhan Shi、Haoyang Li、Wei Bai、Caleb Chen Cao 和 Lei Chen。《显著性方法定量评估的实验研究》。载于 Feida Zhu、Beng Chin Ooi 和 Chunyan Miao（编），《KDD '21：第27届 ACM SIGKDD 知识发现与数据挖掘会议》，线上会议，新加坡，2021年8月14日至18日，第3200–3208页。ACM，2021年。[Miller， 2019] Tim Miller。《人工智能中的解释：来自社会科学的洞见》。《人工智能》，267:1–38，2019年。[Okuyama， 2018] Teruhiro Okuyama。《海马体中的社会记忆印迹》。《神经科学研究》，129:17–23，2018年。日本神经科学奖获奖者的成就。[Quiroga 等人， 2005] R Quian Quiroga、Leila Reddy、Gabriel Kreiman、Christof Koch 和 Itzhak Fried。《人脑中单个神经元形成的不变视觉表征》。《自然》，435(7045):1102–1107，2005年。[Quiroga， 2012] Rodrigo Quian Quiroga。《概念细胞：陈述性记忆功能的构建模块》。《自然综述：神经科学》，13(8):587–597，2012年。[Reddy 和 Thorpe， 2014] Leila Reddy 和 Simon J. Thorpe。《通过联想学习高层表征形成概念细胞》。《神经元》，84(2):248–251，2014年。

[Russakovsky等人， 2015] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、ZhihengHuang、Andrej Karpathy、AdityaKhosla、Michael Bernstein、Alexander C. Berg 和 Li Fei-Fei。《ImageNet 大规模视觉识别挑战》。《国际计算机视觉杂志（*IJCV*）》，115(3):211–252，2015年。

[Sandler等，2018] Mark Sandler、Andrew G. Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen。《MobileNetV2：倒置残差与线性瓶颈》。发表于 *2018*年 *IEEE*计算机视觉与模式识别会议（*CVPR 2018*），美国犹他州盐湖城，*2018*年*6*月*18*日至*22*日，第 4510–4520 页。计算机视觉基金会 / IEEE 计算机学会，2018年。

[Shahid等，2019] Nida Shahid、Tim Rappon 和 Whitney Berta。《人工神经网络在医疗保健组织决策中的应用：范围综述》。《PLOS ONE》，14(2):1–22，2019年02月。

[Shen 和 Huang，2020] Hua Shen 和 Ting-Hao (Kenneth) Huang。《机器生成的解释对普通用户有多大用处？通过人工评估猜测错误预测的标签》。发表于 Lora Aroyo 和 Elena Simperl 编，《第八届 AAAI人类计算与众包会议论文集，HCOMP 2020，荷兰希尔弗瑟姆（线上），2020年10月25日至29日》，第 168–172 页。AAAI Press，2020年。

[Stepin等， 2021] Ilia Stepin、Jos´e Maria Alonso、Alejan- dro Catal´a 和 Martin Pereira-Fari˜na。《面向可解释人工智能的对比式与反事实解释生成方法综述》。《IEEE Access》，9:11974–12001，2021年。

[Tucker等， 2021] Mycal Tucker、Peng Qian 和 Roger Levy。《如果修改了这个，会怎样？利用反事实嵌入进行句法干预》。收录于 Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli 编，《计算语言学协会：*ACL/IJCNLP 2021* 论文集，线上活动，*2021*年*8*月*1*日至*6*日》，ACL/IJCNLP 2021 卷，*ACL* 论文集，第 862–875 页。计算语言学协会，2021年。

[Yang等， 2021] Fan Yang、Ninghao Liu、Mengnan Du 和 Xia Hu。通过属性引导扰动为神经网络生成生成式反事实。*SIGKDD Explor.*，23(1)：59–68，2021年。