arXiv:2304.02202v1 [cs.CV] 5 Apr 2023

借助热力图描述生成和大型语言模型实现深度神经网络的自我解释

Osman Tursun、Simon Denman、Sridha Sridharan 和 Clinton Fookes

SAIVT实验室，昆士兰科技大学，澳大利亚{
osman.tursun,s.denman,s.sridharan,c.fookes}@qut.edu.au

摘要

热力图被广泛用于解释深度神经网络，尤其是在计算机视觉任务中；基于热力图的可解释人工智能（*XAI*）技术也已得到广泛研究。然而，大多数研究都集中于提升生成的热力图质量或探索其他热力图生成技术，而鲜有工作致力于使基于热力图的可解释人工智能自动化、交互化、可扩展且易于使用。为弥补这一不足，我们提出了一个包含两个模块的框架：（*1*）上下文建模；（*2*）推理。我们提出了一种基于模板的图像描述生成方法，用于上下文建模，从热力图和输入数据中生成文本形式的上下文信息。推理模块利用大语言模型，结合专业知识提供解释。定性实验表明，我们的框架和热力图描述生成方法行之有效。我们提出的基于模板的热力图描述方法的代码将公开发布。

1. 引言

深度神经网络在各种机器学习和计算机视觉应用中持续取得非常出色的成果。然而，它们通常被称为“黑箱”模型，因为其决策过程缺乏透明度和可解释性；尽管人们已多次尝试解释这些模型背后的隐藏行为。

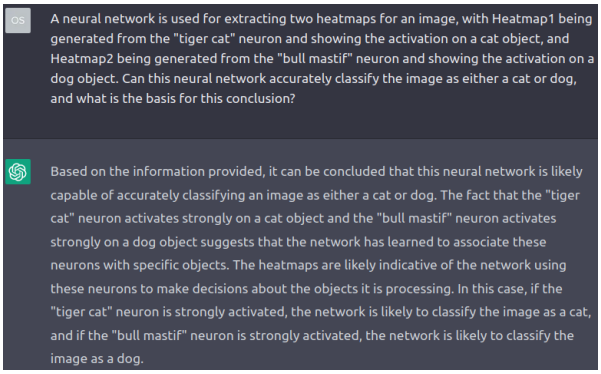
一种受到广泛研究的解释方法是生成热力图，以探查神经网络的决策。热力图是一种图形表示，用于突出显示神经网络在作出最终决策时所关注的数据重要元素。热力图有助于深入了解神经网络，许多研究致力于提高其准确性并改善其质量——



热力图由以下方式生成：可视化“**虎斑猫**”神经元。热力图位于物体上，**猫**

热力图是通过将“**斗牛獒**”神经元可视化生成的。热力图位于物体上，**狗**

（a）热力图以及狗和猫的说明文字



（b）由 ChatGPT 生成的可解释人工智能报告。生成的热力图说明文字会作为提示词发送给 ChatGPT。

图 1. 基于热力图的自解释可解释人工智能所提出框架的示例。在此示例中，ChatGPT 用于生成可解释人工智能报告。

热力图可视化的性质 [16, 18, 21]。然而，准确解读热力图需要上下文知识以及对深度神经网络本身的专业知识；缺少这些信息，任何解读都可能不完整或产生误导。此外，终端用户的可访问性、可扩展性以及热力图解读的自动化

都存在局限。因此，本文聚焦于如何使基于热力图的可解释人工智能（可解释人工智能）实现自动化、可扩展、可解释和交互式，从而让机器学习模型在各种应用中更加可靠、透明，也便于缺乏专业领域知识的用户使用。

我们的方法将图像描述生成与 GPT-3 [2],等大型语言模型的强大能力相结合，使我们能够提取与任务相关的关键信息，并轻松获取专业的专家知识。热力图描述生成以及与大语言模型的交互都可以完全自动化，从而使整个流程易于扩展。我们还可以根据大语言模型的回复，通过用户提示词实现交互式流程。例如，如图1所示，热力图说明会发送给 ChatGPT [1]；ChatGPT 基于 GPT-3 开发。5 [11]在此示例中，我们请 ChatGPT 解释神经网络的能力，并评论它是否能够区分狗类和猫类。聊天机器人会生成一份报告，其中结合了专家知识和本例的具体上下文，而且整个流程完全自动化。根据需求，可以提出各种一般性和技术性问题。这使得基于热力图的可解释人工智能更易于使用，具备交互性、自动化能力和可扩展性。

本研究的主要目标是为热力图生成有意义的说明文字，这不仅是我们所提出框架的第一步，也是最具挑战性的部分。尽管近年来图像描述生成取得了显著进展 [1, 20], , 但大多数方法都是通过自然图像与文本配对数据进行监督训练的。据我们所知，目前还没有适用于热力图的此类数据集。创建合成的热力图与文本配对数据是一种潜在方案，但为热力图生成合成说明文字并非易事，因为热力图在图像中的形状和位置并不规则。此外，热力图说明还需同时考虑源图像和热力图图像。因此，现有方法无法直接应用于热力图描述生成。

为了解决上述问题，我们提出了一种简单的基于模板的方法。该方法利用热力图定位图像中的重要区域，然后提取关键属性以生成说明文字。该方法具有很强的可扩展性，因为可以添加新的属性，从热力图及其源图像中提取更多信息。

总体而言，本工作的贡献总结如下：(i) 我们提出了一种用于基于热力图的自解释型可解释人工智能的全自动框架。该框架使可解释人工智能实现自动化、可扩展且易于使用。(ii) 我们提出了一种用于热力图描述的基于模板的方法，有助于从热力图中提取基于文本的上下文信息。(iii) 通过定性实验，我们展示了该框架和方法的潜力。

[1]chat.openai.com

2. 相关研究

在本研究中，我们主要关注面向神经网络等黑箱模型的基于热力图的可解释人工智能方法。自 AlexNet 获得成功以来，这类方法日益流行，人们也提出了许多不同技术来生成热力图，以解释神经网络的行为。这些技术通常分为三类：基于梯度的方法 [15, 17],基于类别激活的方法 [14,22], 和基于扰动的方法 [3,4]。

近年来，随着先进技术的发展，热力图生成的效率和质量都显著提升 [12, 16, 18, 19, 21]。尽管取得了这些进展，要全面理解热力图并结合给定的任务上下文，仍需专家解读。正如 Kim 等人. [7]所讨论的，基于热力图的解释对具备较高人工智能背景的终端用户而言直观易懂，但对具备较低人工智能背景的终端用户而言则不够直观。热力图说明和易于获取的专业知识可以缩小高、低人工智能背景终端用户之间的差距，使这类方法更易于作为解读方法使用。

人在回路中的可解释人工智能是一个新兴研究领域，专注于将人工反馈融入可解释人工智能流程。所提出的框架支持人工反馈和交互，使可解释人工智能更易于接触、理解和交互。为彰显人工反馈在可解释人工智能中的重要性，Kim 等人 [6] 提出了一种以人为中心的评估框架，在评估流程中考虑用户视角。类似地，Lai 等人 [8]提出了一种利用人工输入生成选择性解释的框架，从而提升人工智能解释的可理解性。这些研究表明，人在回路中的可解释人工智能有望提升可解释人工智能系统的有效性和可用性。

3. 方法

在这项工作中，我们提出了一种全新的框架，能够实现基于热力图的可解释人工智能（XAI）的可扩展性、自我可解释性和交互性。如图所示。2，该框架由两个主要模块组成：上下文和推理。

上下文模块采用所提出的热力图描述生成方法，从输入图像及其对应的热力图中提取任务特定信息。推理模块利用大型语言模型，将这些上下文信息与专业领域知识相结合进行分析。该过程可以完全自动化，因此具有很高的可扩展性；同时，也可以通过引入用户反馈实现交互式处理。为简便起见，我们采用 Chat-

GPT——最先进的基于大语言模型的对话机器人——进行推理，而对于上下文模块，我们将采用下一节所述的方法。

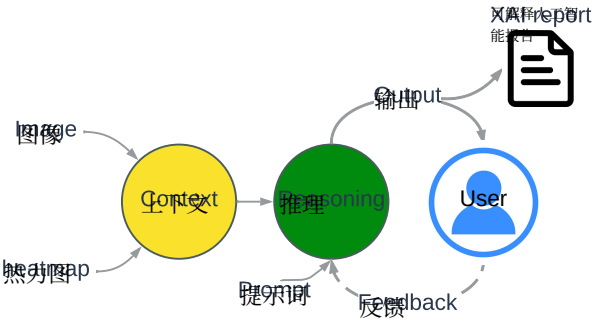


图 2. 所提出的框架示意图。该框架由两个主要模块组成：上下文 和推理。上下文模块通过为热力图生成图注来提供上下文信息。推理模块利用大语言模型，结合专业知识，对上下文信息和用户反馈进行分析。

3.1. 使用基于模板的方法为热力图生成图注

为了提取上下文信息，为热力图图像生成图注至关重要。因此，我们提出一种基于模板的方法来生成热力图的说明文字。与主流的基于深度学习的端到端图像说明文字生成方法相比，基于模板的方法无需在图文对上进行监督训练。基于模板的方法的另一项优势是可扩展性。如有需要，还可以添加其他属性。如图 3所示，我们从给定热力图中所含物体提取了四种属性。下面简要介绍这些属性以及提取步骤。

为提取这四种属性，将使用图像 I 及其灰度热力图 H 。本文选择 SESS[18] 与 GradCAM [14] 来提取热力图。H 用于定位物体和显著区域，而 I 则用于识别和提取细节。通过阈值处理 H 可定位 H 下的所有物体，并应用连通分量分析。每个连通区域均视为一个物体，其矩形边界框则视为该物体的边界框。这里，符号 $(x_i,\ y_i,\ w_i,\ h_i)$ 表示物体 i的矩形边界框。对于每个定位到的物体，将提取以下四种属性。

第一个属性是对象身份。为识别物体，使用 CLIP [13]，语言-图像模型，因为它具有较强的零样本分类能力。对于物体i，裁剪后的区域 $I[x_i:\ x_i+w_i, y_i:\ y_i+h_i]$ 会被发送

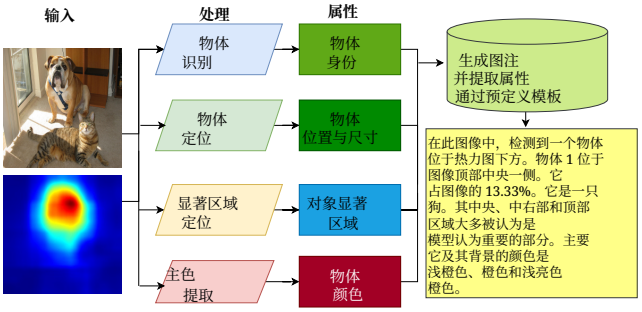


图 3. 所提出的基于模板的热力图图注生成方法示意图，该方法结合四种不同属性来生成热力图图注。

到 CLIP 进行分类。本研究使用 ViT-B/16 Transformer 模型和 COCO 分类标签 [9] 进行分类。

第二个属性是物体的全局位置和尺寸。使用物体边界框的中心来定位物体。 I 被平均划分为九个区域：左上、左中、左下、上中、中心、下中、右上、右中和右下。物体中心（$x_i,\ y_i$）所在的区域被视为物体在图像中的位置。此外，还纳入了物体相对于图像的相对尺寸，其值等于（$w_i * h_i$）$/size$（I）。

第三个属性是物体的显著区域，我们旨在找出对模型决策贡献最大的物体重要区域。对于物体 i，$H[x_i:\ x_i+w_i,\ y_i:\ y_i+h_i]$ 按照全局位置属性的处理方式，被平均划分为九个区域。平均强度值最高的三个区域被视为最显著区域。

物体的最后一个属性指定其主导颜色。该属性通过一种颜色命名算法确定，该算法应用于 HSV 颜色空间，能够用 93 种语义颜色名称之一为像素命名，如 [10] 中所述。为了识别物体 i 的主导颜色，算法根据热图强度值，从区域 $I[x_i$ 中选取前景像素：x_i+w_i，$y_i:\ y_i+h_i]$ 。当像素强度超过 0.5 时，该像素即属于前景。随后，算法为每个选中的像素分配一个颜色名称，并计算每种颜色所占的百分比。百分比最高的三种颜色随后被确定为该物体的主导颜色。

归因提取过程完成后，会将得到的属性填入预先设定的模板中，以生成最终描述，如图3所示。

4. 实验

本节展示了一些定性结果，包括生成的说明文字和使用 ChatGPT 生成的可解释人工智能报告。

在图 4中，我们展示了使用所提出的基于模板的热力图图注方法生成的说明文字示例。这些热力图提取自 ResNet50 [5]。所提出的方法能够为单目标场景（a、b、c）和多目标场景（d）生成说明文字，涵盖全部四项关键属性。生成的说明文字会随着热力图的变化而改变，如（a-b）所示。不过，在某些特定情况下，生成的说明文字并不完全准确。例如，在（c）中，热力图同时覆盖了“人类驾驶员”和“卡丁车”，但图注只包含一个被识别为“汽车”的物体。类似地，在（d）中，一个物体被标记为“风筝”，而另一个与其相同的物体却被标记为“鸟”。这些错误源于基于热力图的物体定位和零样本分类的局限性；采用最先进的目标定位和分类技术可以减少这些错误。

4.1. 使用 ChatGPT 生成可解释人工智能报告

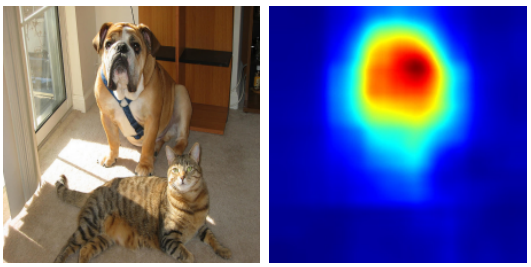
为了检验生成的图像描述是否足够丰富，能否借助 ChatGPT 等大型语言模型生成有意义的可解释人工智能报告，我们将生成的图像描述发送给 ChatGPT，并生成了一些报告。我们要求 ChatGPT 回答以下问题：(i) 模型是否正常工作？(ii) 模型可能存在哪些不足？(iii) 分类模型能否定位某些类型的物体？

提供正确的图像描述和格式规范的提示词后，ChatGPT 会生成措辞良好、内容丰富的可解释人工智能报告。此类报告的示例见补充材料。向 ChatGPT 提供的提示词包含以下内容：

- 热力图的生成方式。例如，用于提取热力图的是哪个神经元。
- 热力图图注。这些图注采用所提出的基于模板的热图描述方法生成，并展示在论文正文的图4中。不过，为了便于从 ChatGPT 获得合理的回答，我们对图4(c-d)中的图注做了少许修改。
- 关于神经网络的问题。我们在以下示例中用下划线标出了这些问题，也用下划线标出了 ChatGPT 对这些问题的回答中的关键部分。

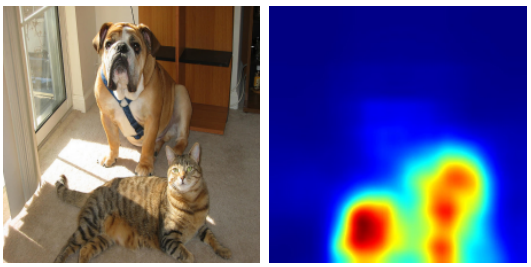
图 4（a-b）的提示词：使用神经网络从图像中提取两张热力图，其中热力图 *1* 是

(a)



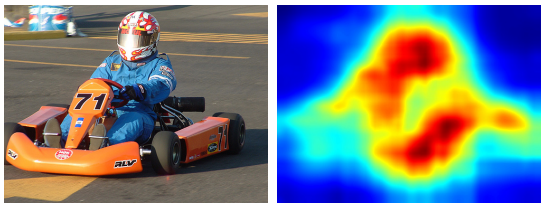
在这张图像中，热力图下检测到一个物体。物体 1 位于图像顶部中央区域，占图像的 13.33%。它是一只狗。模型主要认为它的中央、中右部和顶部中央区域很重要。它及其背景的主要颜色是浅橙色、橙色和浅亮橙色。

(b)



在这张图像中，热力图下检测到一个物体。物体 1 位于图像中下部，占图像的 23.79%。它是一只猫。模型主要认为它的左下部、左中部和中右部区域很重要。它及其背景的主要颜色是浅橙色、橙色和浅黄色。

(c)



在这张图像中，热力图下检测到一个物体。物体 1 位于图像中央一侧，占图像的 68.44%。它是一辆汽车。模型认为其顶部中央、底部中央和中央部分大多较为重要。它及其背景的主要颜色为浅黄色、浅橙色和黑色。

(d)



在这张图像中，热力图下检测到两个物体。物体 1 位于图像左中部，占图像的 12.28%。它是一只风筝。模型主要认为它的中央、中右部和顶部中央区域很重要。它及其背景的主要颜色是蓝色、浅黄色和灰色。物体 2 位于图像中右部，占图像的 8.93%。它是一只鸟。模型主要认为它的中央、中右部和顶部中央区域很重要。它及其背景的主要颜色是蓝色、浅黄色和白色。

图4. 生成的图像描述的定性结果。在每个示例中，都会根据图像及其热力图生成一条图注。

由“虎斑猫”神经元生成，并显示给定图像中猫这一物体上的激活；热力图 *2* 则由“斗牛獒”神经元生成，并显示给定图像中狗这一物体上的激活。以下是热力图的详细信息：热力图 *1*：在此图像中，热力图下检测到一个物体。物体 *1* 位于图像顶部中央，占图像的 *13.33%*。它是一只狗。模型主要认为它的中央、中右部和顶部中央区域较为重要。它及其背景的主要颜色为浅橙色、橙色和浅亮橙色。热力图 *2*：在此图像中，热力图下检测到一个物体。物体 *1* 位于图像中下部，占图像的 *23.79%*。它是一只猫。模型主要认为它的左下部、中左部和中右部区域较为重要。它及其背景的主要颜色为浅橙色、橙色和浅黄色。该神经网络能否准确地将图像分类为猫或狗？这一结论依据是什么？

ChatGPT：根据提供的信息，神经网络生成了两张热力图，一张用于检测图像中的猫，另一张用于检测狗。热力图 *1* 的描述显示，检测到的物体是一只狗，位于图像顶部中央，占图像面积的 *13.33%*。网络判定，狗的中央、中右部和顶部中央区域对检测最为重要；物体及其背景的主要颜色是浅橙色、橙色和浅亮橙色。另一方面，热力图 *2* 的描述显示，检测到的物体是一只猫，位于图像的底部中间，占图像面积的 *23.79%*。网络判定，猫的左下部、中左部和中右部区域对检测最为重要；物体及其背景的主要颜色是浅橙色、橙色和浅黄色。由于网络针对图像中的猫和狗分别生成了热力图，且每张热力图都显示相应物体具有较高的激活程度，因此该神经网络很可能能够准确地将图像分类为猫或狗。不过，若不了解该神经网络的训练数据、架构和性能，就很难对其准确性得出明确结论。网络可能难以区分外观相似的物体，也可能无法很好地泛化到其他图像。要更准确地确定该神经网络的分类准确率，还需要进行进一步的评估和测试。

图4 (c) 的提示词：神经网络将一张图像分类为“卡丁车”，并通过可视化其激活程度最高的神经元生成热力图。以下是对这张热力图的描述：“在这张图像中，热力图下方检测到一个物体。物体 *1* 位于图像中央一侧，占图像的 *68.44%*。它是一辆由人类驾驶员驾驶的卡丁车。模型主要认为它的顶部中央、底部中央和中央部分很重要。它及其背景的主要颜色是浅黄色、浅橙色和黑色。”

这张神经网络可能有什么缺陷？提示：人类驾驶员和卡丁车物体的显著性程度相同。

ChatGPT：这种神经网络的一个潜在缺点是，它没有充分关注那些能够区分人类驾驶员与其他物体或个体的特征。模型对卡丁车和人类驾驶员赋予相同程度的显著性，这表明网络可能没有考虑会影响其分类的重要细节或特征。

例如，如果网络是在不包含人类驾驶员的卡丁车数据集上训练的，它可能没有学会区分有驾驶员和无驾驶员的卡丁车。另一种可能是，网络的架构或训练数据并未针对准确检测和分类人类驾驶员进行优化。

无论是哪种情况，对区分性特征关注不足都可能是这种神经网络的一个潜在缺点，因为这可能导致模型应用于其他图像时出现错误分类或错误预测。为解决这一问题，可以使用更多样化的数据集训练网络，其中包含有驾驶员和无驾驶员的卡丁车图像，以及其他类型的车辆和物体，以确保网络能够准确地区分它们。此外，可能还需要调整网络的架构或训练数据，以提升其识别和区分相关特征的能力。

图4（d）的提示词：使用神经网络提取图像的热力图。热力图的详细信息：在这张图像中，热力图检测到两个物体。物体 *1* 位于图像左中部，占图像的 *12.28%*。它是一只鸟。模型主要认为它的中央、中右部和顶部中央区域很重要。它及其背景的主要颜色为蓝色、淡黄色和灰色。物体 *2* 位于图像中右部，占图像的 *8.93%*。它是一只鸟。模型主要认为它的中央、中右部和顶部中央区域很重要。它及其背景的主要颜色为蓝色、淡黄色和白色。根据热力图信息，这个网络是否适合定位鸟类物体？

ChatGPT：根据所提供的信息，该神经网络很可能能够准确检测并定位图像中的鸟类目标。网络生成的热力图表明，图像中检测到了两个鸟类目标，它们分别占据图像中的不同区域，并由若干重要区域识别出来，包括中央、中右部和顶部中央。

每个目标。此外，对各目标及其背景主要颜色的描述表明，该网络能够识别并区分鸟类目标相关的不同颜色模式。总体而言，所提供的热力图信息表明，该神经网络有助于定位图像中的鸟类目标，尤其是那些具有明显颜色和特征模式、易于被网络区分的目标。不过，需要注意的是，网络的准确性和有效性可能会因训练数据的质量与多样性，以及网络架构的复杂性和适应能力而有所不同。

5. 未来工作与结论

在本研究中，我们提出了一个利用大型语言模型实现自动化、交互式且可扩展的可解释人工智能报告生成的前景良好的框架。具体而言，我们采用了所提出的基于模板的热力图描述生成方法，为 ChatGPT 等基于大型语言模型的推理模块提供上下文信息。结果表明，该方法具有良好的应用前景，同时也凸显了准确的基于模板的描述生成方法的重要性；我们注意到，要实现完全自动化的可解释人工智能报告生成，仍需进一步改进。我们观察到，尽管生成的说明文字包含丰富信息，但缺乏多样性，并且可能包含冗余信息。未来的研究可以探索使用基于深度学习的图像描述生成方法，以解决这些局限。此外，我们发现 ChatGPT 生成的报告不够简洁，并且需要经过精心设计的提示词。尽管存在这些不足，专门针对可解释人工智能报告生成设计大型语言模型仍展现出良好的研究前景。

参考文献

[1] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katie Millican、Malcolm Reynolds 等。Flamingo：用于小样本学习的视觉语言模型。*arXiv* 预印本 *arXiv:2204.14198*，2022年。2[2] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是小样本学习器。神经信息处理系统进展，33：1877–1901，2020年。2[3] Piotr Dabkowski 和 Yarin Gal。黑盒分类器的实时图像显著性。神经信息处理系统进展，30，2017年。2[4] Ruth C Fong 和 Andrea Vedaldi。通过有意义的扰动对黑盒模型进行可解释说明。在*IEEE*国际计算机视觉会议论文集，第3429–3437页，2017年。2

[5] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于《IEEE计算机视觉与模式识别会议论文集》，第770–778页，2016年。4[6] Sunnie SY Kim、Nicole Meister、Vikram V Ramaswamy、Ruth Fong 和 Olga Russakovsky。Hive：评估视觉解释的人类可解释性。载于《计算机视觉—ECCV 2022：第17届欧洲计算机视觉会议，特拉维夫，以色列，2022年10月23日至27日，论文集，第XII部分》，第280–298页。Springer，2022年。2[7] Sunnie SY Kim、Elizabeth Anne Watkins、Olga Russakovsky、Ruth Fong 和 Andr´es Monroy-Hern´andez。“帮我帮助人工智能”：理解可解释性如何支持人机交互。arXiv 预印本 arXiv:2210 03735，2022年 2 . .[8] Vivian Lai、Yiming Zhang、Chacha Chen、Q Vera Liao 和 Chenhao Tan。选择性解释：利用人工输入使可解释人工智能保持一致。arXiv预印本 arXiv:2301 09656，2023年 2 . .[9] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于《欧洲计算机视觉会议论文集》，第740–755页。Springer，2014年。3[10] Ying Liu、Dengsheng Zhang、Guojun Lu 和 Wei-Ying Ma。利用高级语义颜色名称进行基于区域的图像检索。载于第11届国际多媒体建模会议论文集，第180–187页。IEEE，2005年。3[11] Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agar- wal、Katarina Slama、Alex Ray 等。训练语言模型，使其能够遵循指令并接受人工反馈。arXiv预印本 arXiv:2203.02155，2022年。2[12] Vitali Petsiuk、Abir Das 和 Kate Saenko。RISE：通过随机输入采样解释黑箱模型。arXiv预印本 arXiv:1806.07421，2018年。2[13]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于《国际机器学习会议论文集》，第8748–8763页。PMLR，2021年。3[14] Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。Grad-CAM：通过基于梯度的定位为深度网络提供视觉解释。载于《IEEE国际计算机视觉会议论文集》，第618 626页，2017年 2，3 – 。[15] Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman。深入卷积网络：图像分类模型与显著图的可视化。arXiv预印本 arXiv:1312.6034，2013年。2[16] Daniel Smilkov、Nikhil Thorat、Been Kim、Fernanda Vi´egas 和 Martin Wattenberg。SmoothGrad：通过添加噪声消除噪声。arXiv 预印本 arXiv:1706.03825，2017年。1，2[17] Mukund Sundararajan、Ankur Taly 和 Qiqi Yan。深度网络的公理化归因。载于《国际机器学习会议论文集》，第3319–3328页。PMLR，2017年。2

[18] Osman Tursun、Simon Denman、Sridha Sridharan 和 Clin- ton Fookes。SESS：通过缩放与滑动增强显著性。载于《计算机视觉—*ECCV 2022*：第*17*届欧洲计算机视觉会议，特拉维夫，以色列，*2022*年*10*月*23*日至*27*日，论文集，第*XII*部分》，第318–333页。Springer，2022年。1，2，3[19] Haofan Wang、Zifan Wang、Mengnan Du、Fan Yang、Zijian Zhang、Sirui Ding、Piotr Mardziel 和 Xia Hu。Score-CAM：针对卷积神经网络的加权视觉解释。载于《*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集》，第24–25页，2020年。2[20] Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel 和 Yoshua Bengio。看、关注并讲述：利用视觉注意力生成神经图像图注。载于《国际机器学习会议》，第2048–2057页。PMLR，2015年。2[21] Qinglong Zhang、Lu Rao 和 Yubin Yang。Group-CAM：针对深度卷积网络的组加权视觉解释。*arXiv*预印本 *arXiv:2103.13859*，2021年。1，2[22] Bolei Zhou、Aditya Khosla、Agata Lapedriza、Aude Oliva 和 Antonio Torralba。用于判别式定位的深度特征学习。载于《*IEEE*计算机视觉与模式识别会议论文集》，第2921 2929页，2016年2– 。