软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(4):1591−1614 [doi: 10.13328/j.cnki.jos.007530] [CSTR: 32375.14.jos.007530]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 可解释深度学习的概念建模方法综述*

王家祺[1,2], 冯 毅[1,2], 刘华锋[1,2], 景丽萍[1,2], 于 剑[1,2]

[1](交通数据挖掘与具身智能北京市重点实验室 (北京交通大学), 北京 100044)

[2](北京交通大学 计算机科学与技术学院, 北京 100044)

通信作者: 刘华锋, E-mail: hfliu1@bjtu.edu.cn; 景丽萍, E-mail: lpjing@bjtu.edu.cn

**摘 要**: 近年来, 深度神经网络在多个领域取得了显著进展, 但其作为典型的黑盒模型, 内部机制仍难以为人所理解, 给医疗诊断、金融风控、自动驾驶等高风险应用场景带来了严峻挑战. 提升模型的可解释性, 已成为实现高可信机器学习的核心问题之一. 现有可解释性方法大致可分为两类: 基于信息流的解释和基于概念的解释. 基于信息流的解释主要侧重于神经元或特征重要性分析, 如定位图片中对分类结果起关键作用的像素区域. 虽然能揭示模型“关注了什么”, 但难以提供具备人类语义的认知解释; 相比之下, 基于概念的解释通过构建语义空间, 将模型内部表示映射为可理解的概念结构, 能够以“模型理解了什么”的方式提供更具语义深度和认知契合的解释, 在增强语义透明性和用户信任方面展现出独特优势. 深度学习的不可解释性源于其语义表达的缺失, 因此, 如何构建对人类认知友好的概念空间与表示机制, 已成为可解释模型研究的关键突破口. 围绕可解释深度学习中的概念建模方法展开综述, 依据建模介入阶段将相关研究划分为事后解释与事中解释两大路径: 前者通过神经元解剖、语义聚类等手段挖掘已有模型的概念表示, 后者则在训练过程中引入结构化先验或语义约束, 以实现模型的内生可解释性. 基于该分类框架, 系统梳理了典型方法的建模思路与代表性成果, 比较其在语义透明性与实际应用中的性能差异, 并总结当前研究面临的挑战与未来发展方向, 旨在为理解和构建语义可解释的深度模型提供系统性参考与方法指引.

**关键词**: 可解释性; 深度学习; 概念表示; 事后解释; 自解释

**中图法分类号**: TP18

中文引用格式: 王家祺, 冯毅, 刘华锋, 景丽萍, 于剑. 可解释深度学习的概念建模方法综述. 软件学报, 2026, 37(4): 1591–1614. http://www.jos.org.cn/1000-9825/7530.htm

英文引用格式: Wang JQ, Feng Y, Liu HF, Jing LP, Yu J. Survey on Concept-based Modeling Methods for Interpretable Deep Learning. Ruan Jian Xue Bao/Journal of Software, 2026, 37(4): 1591–1614 (in Chinese). http://www.jos.org.cn/1000-9825/7530.htm

## Survey on Concept-based Modeling Methods for Interpretable Deep Learning

WANG Jia-Qi[1,2], FENG Yi[1,2], LIU Hua-Feng[1,2], JING Li-Ping[1,2], YU Jian[1,2]

[1](Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence (Beijing Jiaotong University), Beijing 100044, China)

[2](School of Computer Science & Technology, Beijing Jiaotong University, Beijing 100044, China)

**Abstract**: In recent years, deep neural networks have achieved significant progress across various domains. However, as typical black-box models, their internal mechanisms remain difficult for humans to understand, posing serious challenges in high-stakes applications such as medical diagnosis, financial risk management, and autonomous driving. Enhancing model interpretability has become one of the core issues in building highly trustworthy machine learning systems. Existing interpretability methods can be broadly classified into two categories:

* 基金项目: 国家重点研发计划 (2024YFE0202900); 国家自然科学基金 (62436001); 国家自然科学基金青年基金 (62406019); 北京市自然科学基金青年基金 (4244096); 北京交通大学人才基金 (2024XKRC075)
王家祺和冯毅为共同第一作者.
本文由“高可信自适应机器学习”专题特约编辑王熙照教授、张敏灵教授、曹付元教授推荐.
收稿时间: 2025-05-12; 修改时间: 2025-06-30, 2025-08-15; 采用时间: 2025-08-20; jos 在线出版时间: 2025-09-02
CNKI 网络首发时间: 2025-12-11

information-flow-based explanations, which focus on analyzing the importance of neurons or features, such as locating the pixel regions in an image that play a decisive role in classification results. Although these methods can reveal what the model has “attended to”, they often fail to provide cognitively meaningful, human-understandable semantics. In contrast, concept-based explanations construct semantic spaces to map internal model representations to interpretable concept structures, thus answering what the model has understood. These methods offer greater semantic depth and cognitive alignment, making them especially effective in improving semantic transparency and user trust. The fundamental lack of interpretability in deep learning stems from its deficiency in semantic representation. Therefore, constructing concept spaces and representation mechanisms aligned with human cognition has become a key breakthrough point in the development of interpretable models. This study presents a comprehensive survey of concept-based modeling methods in interpretable deep learning. Based on the stage at which interpretability is introduced, existing approaches are categorized into two major paradigms: post-hoc explanations, which extract semantic representations from trained models through techniques such as neuron dissection and semantic clustering; and intrinsic explanations, which incorporate structured priors or semantic constraints during training to endow models with built-in interpretability. Within this classification framework, this study systematically reviews representative modeling strategies and key methods, compares their performance in terms of semantic transparency and practical applicability, and summarizes current challenges and future research directions. The goal is to provide a structured reference and methodological guidance for understanding and building semantically interpretable deep learning models.

**Key words**: interpretability; deep learning (DL); conceptual representation; post-hoc explanation; intrinsic explanation

近年来, 深度神经网络凭借其强大的特征自动提取与建模能力, 在图像识别[1−4]、自然语言处理[5]、推荐系统[6]、自动驾驶[7]等多个领域取得了显著进展, 成为推动人工智能快速演进的核心技术之一. 然而, 深度模型在实现高性能的同时, 也表现出高度的不透明性, 其复杂的内部机制难以为人所理解, 被广泛视为“黑盒”模型. 这种不可解释性在某些通用任务中可能尚可接受, 但在医疗诊断[8,9]、金融风控[10,11]、自动驾驶[12,13]等安全性和可靠性要求极高的场景中, 模型一旦出现异常或错误决策, 可能带来严重后果. 因此, 如何增强深度模型的可解释性, 已成为实现高可信机器学习的核心技术瓶颈与研究热点[14,15].

目前, 可解释人工智能 (explainable artificial intelligence, XAI) 已形成初步体系, 研究者提出了多种方法[16−18]以揭示深度模型的行为逻辑. 总体来看, 现有可解释性方法大致可分为两类路径: 基于信息流的解释方法与基于概念的解释方法. 前者通过分析模型内部的信息传递过程 (如梯度传播、特征响应、注意力分布等), 尝试定位模型在输入样本中“关注了哪些特征”, 揭示其决策依据. 典型方法包括敏感度分析 (sensitivity analysis)[19,20]、逐层相关性传播 (layer-wise relevance propagation, LRP)[21]、梯度加权类激活映射 (Grad-CAM)[22]等, 能够以热力图等形式直观展示模型聚焦区域. 尽管这些方法在可视化和模型调试方面具有一定实用性, 但其解释往往停留在底层像素或局部特征层面, 缺乏与人类认知一致的语义结构支撑, 难以提供具有可理解性的概念性解释.

相比之下, 基于概念的解释方法强调引入具有人类语义的中间表示, 尝试回答“模型理解了什么”这一更具认知价值的问题. 该类方法通过构建或挖掘语义概念空间, 将模型内部特征与人类可感知、可命名的语义单位建立映射, 从而使模型决策过程具备更强的透明性和可解释性. 例如, 概念激活向量测试 (testing with concept activation vectors, TCAV)[23]方法通过定义一组正负样本生成方向向量, 解释模型是否在使用某个概念; 原型网络[24]通过嵌入语义原型增强模型结构可解释性; 而基于聚类分析的语义模式提取方法则尝试在高维特征空间中发现潜在概念簇. 相比于低层特征可视化, 概念级解释可直接对应人类理解的语义单元, 在提升模型可控性、支持审计追责和辅助领域专家决策等方面具有显著优势.

从深层次看, 深度学习的不可解释性根源在于其内部表示缺乏语义清晰的认知结构[25]. 在人类认知中, 概念是构建语义理解的基本单元, 是人们识别、分类和推理的核心载体. 因此, 若想让深度模型具备可解释性, 关键在于使其内部表征能够映射为人类可理解的概念空间. 在早期的哲学与逻辑传统中, 概念被认为可以通过一组清晰的符号、完备的规则或命题进行定义, 这种观点被称为经典内涵理论[26]. 根据这一理论, 某个对象是否属于某一概念, 是一个严格的二值判断问题, 只需验证其是否满足人为设定的必要与充分条件. 例如, 在交通法规中, “限速标志”可以通过其几何形状和文字内容进行明确定义; 在程序语言中, “变量名”必须符合特定语法规则. 这些概念具有清晰的边界和逻辑定义, 因此适用于经典内涵的形式化表示.

然而, 现实世界中的许多概念并不具备如此严格和统一的定义边界. 以图像分类任务为例, 某些常见类别, 如

“厨房”“SUV 车型”或“奔跑的动物”, 往往因拍摄角度、背景环境、遮挡因素、类内变异等因素而呈现出显著的多样性. “厨房”可能包含橱柜、灶台、水槽、微波炉等多种组合形式, 其判定标准难以归结为固定规则; “奔跑的动物”不仅涉及物体本身的外观, 还涉及动态姿态和语义情境. 这类概念的边界具有明显的模糊性和上下文依赖性, 难以通过一组稳定命题进行精确定义. 这些认知上的模糊与复杂暴露了经典内涵理论的局限性. 随着认知科学的发展, 人们逐渐认识到, 大量现实概念更符合原型结构、样例积累和知识依赖的认知模型. 这也促使人工智能研究引入现代认知视角, 尝试通过原型表示、相似模式聚类、语言描述等方式建模概念的现代内涵, 以增强模型的语义可解释性和人机认知一致性.

有别于此前对深度学习可解释性方法的综述工作, 本文聚焦可解释深度学习中的概念建模方法, 系统总结其在建模策略、解释机制与应用实践的研究进展和技术路径. 以模型解释机制介入的时间点为划分依据, 将现有工作系统性地归纳为两大路径: 事后解释方法与事中解释方法. 前者通过模型分析和语义挖掘, 在训练完成后重构模型内部的语义结构; 后者则在模型设计和训练过程中, 主动嵌入可解释结构、先验知识或语义约束, 从而实现内生的解释能力. 在此分类框架下, 本文将梳理不同方法的建模思路与技术特点, 比较其在语义透明性、应用适应性与性能保障方面的差异, 并分析当前研究所面临的主要挑战与发展趋势, 旨在为构建语义层面可解释的深度模型提供系统参考与理论支撑.

本文第 1 节对可解释人工智能的基本定义及其在深度学习中的研究背景进行综述, 重点梳理“可解释性”与“概念”的关系, 并基于认知科学视角对概念的内涵与外延进行分类, 明确“概念建模”在可解释性研究中的理论基础与表达方式. 第 2 节对已有的事后可解释方法中概念建模的相关研究进行总结, 涵盖激活可视化、神经元解剖、概念向量分析等方法, 并对其优势与局限进行分析. 第 3 节对事中可解释方法中的概念建模策略进行系统归纳, 涉及原型网络、概念瓶颈网络以及相关可解释性损失函数等机制, 探讨模型内生可解释性的实现路径. 第 4 节总结可解释性研究中常用的评价指标与度量手段, 包括语义一致性、人类认知对齐等方面的评估方法, 并评述当前在概念层面可解释性度量方面的挑战. 第 5 节对该领域未来值得关注的研究方向进行展望, 特别关注概念抽象层次的泛化能力、跨模态语义一致性以及与大模型语义表示的融合趋势等问题.

## 1 可解释深度学习

### 1.1 可解释性的基本定义

在可解释人工智能领域, 英文术语“Explanation”与“Interpretation”虽在中文中常被统称为“可解释”, 但在英文语境下二者存在细微而重要的区别. 理解这一差异, 有助于厘清不同可解释性方法的侧重点及技术路径.

根据文献 [19] 的定义, Interpretation 通常指的是将模型内部的抽象表示 (如某个类别、语义标签或中间预测结果) 映射到一个人类能够理解的认知域中, 即解释整个模型或其行为规律的过程. 例如, 某深度分类模型预测图像属于“鸟类”这一类别时, 若能提供模型对“鸟类”概念的整体理解方式, 如对羽毛形状、颜色或身体结构的关注模式, 则属于解释的整体性建模范畴. 而 Explanation 更强调的是对单个预测决策的具体成因进行剖析, 即在某一特定输入 (如一张鸟类图片) 下, 模型依赖了哪些特征做出该决策. 它通常被定义为: “在可解释特征空间中, 对影响模型某一决策结果的要素集合的识别过程.” 这种解释往往以热力图、特征掩码、梯度权重等形式呈现, 用以回答“模型是基于哪些像素/特征做出该判断”的问题. 本文在此基础上总结出两类主流的可解释性方法.

一类是基于信息流的解释方法, 主要通过分析模型内部特征流动和传播路径来定位重要输入区域. 该类方法的核心思想是: 通过分析模型内部信息如何在各层之间流动, 以及特征在输入与输出之间如何被利用, 从而估计哪些输入区域或特征对模型预测起到了关键作用. 这类方法通常回答的是“模型关注了什么”, 其目标是标记出输入中最具决策影响力的部分. 经典的信息流解释方法可大致分为前向扰动方法与后向传播方法. 前向扰动方法 (forward-perturbation method)[27]通过对输入进行遮蔽、替换或模糊等有控制的干扰, 观察输出变化来评估输入各区域的重要性. 例如, 可遮挡输入图像的某个区域, 并测量模型输出的波动程度, 从而判断该区域对决策的影响. 该类方法的优点是实现简单、解释直观、不依赖模型梯度信息, 因此适用于各种模型结构; 但由于需要对每个区域

重复进行前向传播, 计算开销较大.

与之相对, 后向传播方法 (back-propagation method) 通过一次或有限次反向传播即可估计不同神经元或输入特征的重要度, 具有更高的效率. 典型方法包括敏感度分析[19,20], 其通过计算模型输出对输入的梯度, 衡量输入扰动对预测的影响程度, 适用于生成显著性图或热力图. 另一代表方法是逐层相关传播[21], 该方法从输出向输入逐层传播“相关性值”, 通过规则化反向传播过程将预测结果分配回输入维度, 从而揭示各输入部分在决策中的具体贡献. 相比直接使用梯度, LRP 同时考虑了输入激活和梯度信息, 在深层网络中具有更强的稳定性和解释鲁棒性.

另一类是基于概念的解释方法 (concept-based explanation method), 其核心目标不再是简单地定位模型“关注了什么”, 而是试图从语义层面回答“模型理解了什么”. 这类方法强调构建与人类认知一致的语义中间表示, 以实现对模型内部特征表示的更高层次解释. 相比信息流方法所提供的像素级显著性图, 基于概念的解释更具语义深度和认知契合性. 具体而言, 概念解释方法通常依赖于引入人类可识别的概念标签 (如“翅膀”“斑马纹”“车窗”)、原型样本 (代表性图像区域或特征块), 或构建语义嵌入空间, 在模型特征与概念语义之间建立映射. 典型方法如 TCAV[23], 通过在高层特征空间中训练一个支持向量机区分概念样本与非概念样本, 并将其分类边界的法向量作为概念方向, 用以衡量模型在预测过程中对特定语义概念的依赖程度. 另一种经典的方法 ProtoPNet (prototypical part network)[24], 在网络中显式引入概念原型模块, 通过对输入样本与若干“原型特征图”的相似度匹配, 实现可追溯的分类决策路径. 每个原型代表一个可视化的语义概念 (如“鸟的头部”“车的轮胎”), 可通过原型-图像的对齐图示明确呈现, 用户可以直观理解模型为何做出当前预测.

此外, 随着多模态模型和大语言模型的发展, 部分研究开始引入自然语言描述作为概念表达形式, 使模型具备生成式的解释能力. 这类方法在语言与图像、概念之间建立对应, 有助于在人机交互或审计任务中提升可用性. 图 1 对比了两种方法, 前者更侧重标注模型关注的输入区域或像素点, 后者则尝试构建“概念组合”作为模型决策依据.

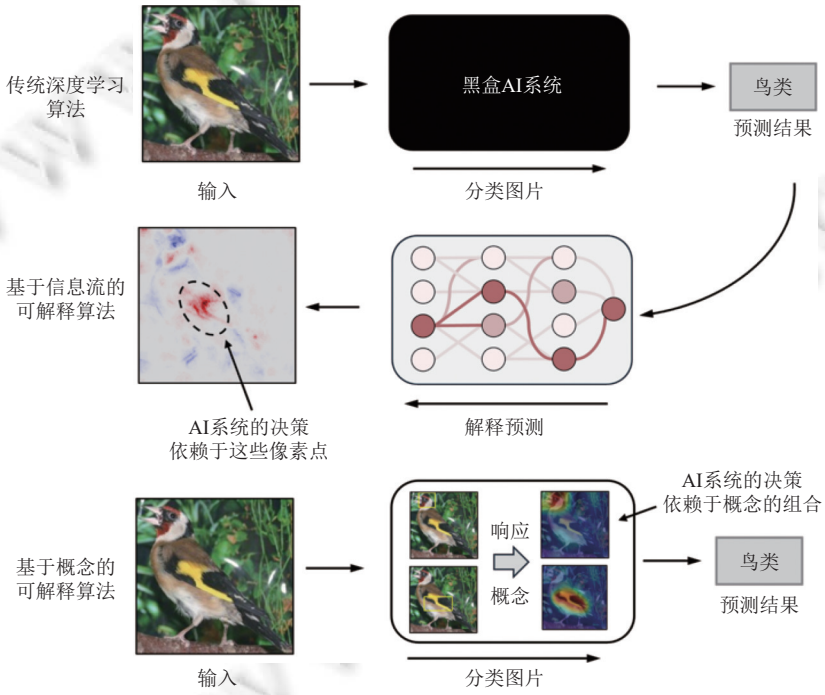


图 1 基于信息流的可解释算法与基于概念的可解释算法

总体而言, 基于信息流的解释方法强调特征层面的可视化与重要性评估, 适用于快速定位模型关注区域, 但难以揭示更高层次的语义关联; 而基于概念的解释方法则通过构建与人类认知相符的语义结构, 为理解模型内部机制提供了更具认知深度的视角. 两类方法在解释粒度、表达方式与可理解性上各具优势, 共同构成了当前可解释

深度学习研究的重要基础. 基于概念的方法在提高模型语义透明性、构建人类理解层的因果链路、增强模型的可控性与信任感等方面展现出独特优势, 尤其适用于需要与人类认知对齐的高风险场景. 正因如此, 概念建模已成为近年来可解释性研究中最受关注和发展最快的分支方向之一.

### 1.2 概念表示形式

在可解释人工智能中, “概念”不仅是人类认知中的基本语义单位, 更是连接深度神经网络内部表示与人类理解之间的关键桥梁. 为了实现对深度模型“理解了什么”的语义层解释, 亟需在机器可计算的表示空间中对“概念”进行有效的构建、表达与操作. 这不仅要求从理论上界定概念的表示形式, 还需结合具体任务中的建模实践, 分析其在不同解释机制下的适配方式与表现能力.

“概念”的定义存在多样性, Molnar[28]曾将其广义描述为“任何抽象实体, 如颜色、物体, 甚至是想法”. 从认知视角来看, 概念是在特定知识背景下, 人类对一类具有共同特征对象的抽象总结, 是感知数据与语义理解之间的重要中介. 概念通常包含内涵与外延两个层面: 内涵 (intension) 指概念本质属性的规则、特征或原型, 是人类在心智中对该概念的认知结构, 常用命题、向量或结构模型表示; 外延 (extension) 则是满足这些内涵属性的具体实例集合, 体现为数据样本或可视图像, 是概念在现实世界中的投射. 综上, 本文对概念的形式化表达如下:

$$C = \{x \in \mathcal{X} | \phi(x) = 1\},$$

其中, $\mathcal{X}$ 表示输入空间 (图像、文本等), $\phi(x)$ 表示概念内涵的认知函数 (如特征组合、神经激活规则等), $C$ 表示概念的外延, 即现实世界中满足内涵的实例集合. 本文将概念内涵归纳为 4 种主要类型: 符号化概念 (symbolic concept)、相似模式簇 (similar pattern cluster)、原型 (prototype) 以及文本概念 (textual concept). 每一种类型都代表了概念在可解释性研究中的独特角色和应用方式.

(1) 符号化概念. 符号化概念是指由人类预先定义并具有明确语义标签的高级抽象概念, 通常用于对模型内部行为进行解释和对齐. 这类概念通常以属性、词语或结构性标签的形式存在, 如“翅膀”“红色”“圆形边缘”等, 能够与特定视觉或语义特征直接对应. 在实际应用中, 符号化概念往往依赖于专家知识、人工注释或辅助元数据, 通过将模型特征与这些语义单位建立映射, 提升可解释性的透明度与认知一致性. 例如, 在鸟类图像分类任务中, “翅膀”或“鸟喙”可以作为符号化概念, 用于说明模型是基于哪些语义因素做出决策. 与其他类型的概念相比, 符号化概念具备高度的可控性和明确性, 既可用于在事后分析中进行语义对齐, 也常用于在事中模型设计中作为先验引导, 是当前可解释性建模中最常用的概念类型之一.

(2) 相似模式簇. 相似模式簇是指神经网络内部激活特征中具有相似空间分布或响应模式的一组聚类区域. 这些聚类往往隐含着某种潜在的高阶语义结构, 例如纹理、边缘形状或局部结构模式, 尽管它们并非由人类直接标注, 但在聚合后往往能对应某种可感知的抽象概念, 如“斑马的条纹”或“建筑的窗格”. 这类概念通常通过无监督聚类算法从网络中间层特征中提取, 可在模型训练阶段或训练完成后进行分析, 因此被广泛应用于事中建模与事后解释两类方法中. 与符号化概念相比, 相似模式簇更强调概念的自发性与可挖掘性, 无需预设标签, 适用于数据驱动的解释机制. 在自动化概念构建与模型内部语义结构重构等任务中, 该类型概念发挥着重要作用, 是实现可扩展性与语义泛化的重要手段之一.

(3) 原型. 原型是指训练数据中某一类样本或其局部区域中最具代表性的实例, 用于表达该类别在语义或特征空间中的核心特征. 原型可以是完整样本 (如一张图像), 也可以是样本中关键的局部区域 (如鸟类图像中的头部区域), 其作用是作为模型决策的可视化依据与语义参考点, 帮助解释“模型为什么将该输入判为某类”. 与无监督生成的相似模式簇不同, 原型通常以结构化方式嵌入到模型中, 作为显式的模型组件进行建模, 例如在 ProtoPNet[24]等原型网络中, 每一个类别都由若干可视化原型构成, 模型通过比较输入与原型之间的距离做出分类决策. 这种方式不仅实现了端到端的可解释性, 也使得模型内部表示具备可视追踪和人类语义对齐的能力. 由于原型需要在模型训练过程中显式编码和优化, 它主要用于事中可解释方法, 不适用于训练完成后的事后分析. 原型方法特别适合对透明性要求较高的任务场景, 如医疗图像诊断.

(4) 文本概念. 文本概念是指通过简洁的自然语言短语对类别或中间语义特征进行描述的概念形式, 旨在以人

类可理解的语言表达模型内部表示的语义含义. 例如, 在鸟类分类任务中, 文本概念可以表现为“一只黑头黄身的小鸟”或“具有长喙和灰色翅膀的水鸟”等, 这些描述语句能够作为模型预测的语义解释依据. 早期的文本概念主要依赖于人工标注或专家先验, 常用于多标签分类、视觉问答等任务中的辅助信息引入. 随着大语言模型 (large language model, LLM) 与多模态预训练模型的发展, 近年来大量研究开始探索使用自动化方式从图像或模型激活中生成文本概念. 这使得文本概念能够以可扩展、低成本的方式嵌入到模型训练或解释框架中, 增强模型的语言解释能力与人机交互性. 文本概念多用于事中可解释模型的构建, 如语言引导的注意力机制、描述驱动的生成解释等, 尤其适用于需要符号层语义输出的高风险场景, 如医疗辅助诊断、司法证据呈现等. 其优势在于解释结果的自然语言可读性与跨模态通用性, 是当前可解释性研究中备受关注的新兴方向.

除了根据概念的表示形式进行划分外, 基于概念的可解释性方法还可按照解释结构与作用机制划分为 3 种典型类型: 类-概念关系 (concept-class, C-C)、神经元-概念关联 (neuron-concept, N-C) 和概念可视化 (visualization, V). 这 3 类方法从不同层面提供了解释的视角, 以适应多样化的应用需求和模型理解目标.

类-概念关系 (C-C): 该类方法关注模型中某一概念与输出类别之间的关系建模, 能够揭示特定概念在分类决策中所起的作用. 例如, 一张图像被分类为“熊猫”, 可能是因为模型检测到了“毛茸茸的耳朵”“黑色眼圈”等多个概念特征. 通过构建概念与类别之间的映射规则, 模型能够更清晰地表达其决策的语义依据, 有助于增强用户对模型行为的信任与可控性.

神经元-概念关联 (N-C): 此类方法旨在将语义概念与神经网络中的特定神经元或通道建立对应关系, 提升模型内部表示的透明度. 其实现方式可以是事后分析神经元的激活响应, 找出与特定概念高度相关的神经元; 也可以在训练阶段引入监督机制, 使模型内部单元显式学习特定概念的表示. 这种方式有助于揭示模型在处理输入时的语义分布结构, 增强对其内部工作机制的认知理解.

概念可视化 (V): 通过热力图、显著性图等可视化手段, 突出输入样本中最能代表某一概念的区域或特征, 从而帮助用户直观地理解模型是如何在输入数据中捕捉与响应语义模式的. 这种可视化不仅提升了模型解释的直观性, 也在模型注意力机制与人类感知之间建立起联通, 增强模型行为的可感知性和可解释性.

此外, 基于概念的可解释方法还可以按照解释的范围划分为全局解释与局部解释. 全局解释侧重于对整个模型的语义结构和行为逻辑进行宏观理解, 例如识别模型使用了哪些核心概念及它们之间的关系; 局部解释关注模型在特定输入样本下的具体决策过程, 通过分析样本触发的概念激活路径, 为用户提供个性化、场景化的解释信息.

图 2 展示了本文中基于概念的可解释方法的分类框架, 帮助用户根据是否需要对模型进行结构修改、使用何种类型的概念表示, 以及期望获取何种粒度的解释效果, 选择最适合的方法.

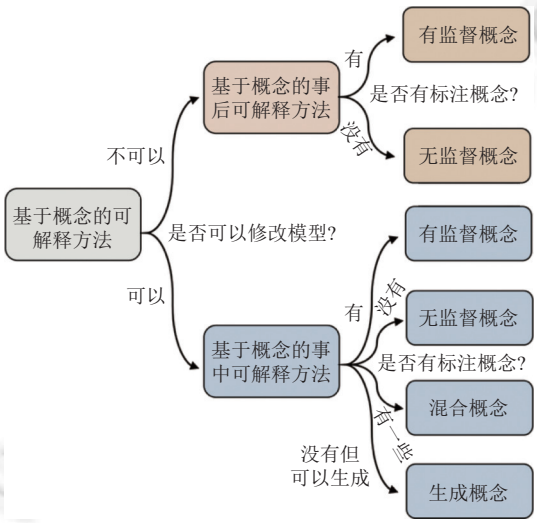


图 2 可解释概念建模方法的分类框架及选择指南

## 2 事后可解释的概念建模方法

基于概念的解释方法允许在不干扰模型内部结构的前提下, 对模型进行深入分析, 因此被称为基于概念的事后可解释方法 (post-hoc interpretable method). 该类方法通常在模型训练完成后实施, 旨在提供关于模型所学习到的概念表示及其决策依据的洞察. 根据是否使用概念标注信息, 事后方法可进一步分为有监督方法与无监督方法. 其中, 有监督方法依赖于带有概念标注的数据, 通过引入明确定义的符号概念来引导模型行为的解释过程; 而无监督方法则通过聚类等技术, 从模型的中间表示中自动识别潜在的概念簇, 这些簇基于数据自身结构而形成, 无需预先设定标签, 从而为模型决策提供了一种更数据驱动、探索性的解释路径.

为系统梳理该类方法, 本文构建了一个多维度分析框架, 从概念类型、解释类型、解释范围、数据类型、任务场景及网络结构等维度对典型的事后概念方法进行了整理与比较 (见表 1). 其中, 网络类型涵盖了卷积神经网络 (CNN)[1−3]和变换器架构 (Transformer, 缩写为 TRANSF)[4,5]等主流模型.

表 1 基于概念的事后解释性方法

| 概念标注 | 方法 | 概念类型 | 解释类型 | 解释范围 | 数据类型 | 网络类型 |
|---|---|---|---|---|---|---|
| 有监督 | TCAV[23] | 符号化概念 | C-C | 局部/全局 | 图像 | CNN |
| | CARs[29] | | C-C, V | 局部/全局 | 图像 | CNN |
| | IBD[30] | | C-C, V | 局部/全局 | 图像 | CNN |
| | CaCE[31] | | C-C | 全局 | 图像 | CNN |
| | CPM[32] | | C-C | 局部 | 文本 | TRANSF |
| | Object detector[33] | | N-C, V | 全局 | 图像 | CNN |
| | ND[34] | | N-C, V | 全局 | 图像 | CNN |
| | Net2Vec[35] | | N-C, V | 全局 | 图像 | CNN |
| 无监督 | ACE[36] | 相似模式簇 | C-C, V | 局部/全局 | 图像 | CNN |
| | Completeness-aware CE[37] | | C-C | 全局 | 图像/文本 | CNN |
| | ICE[38] | | C-C, V | 局部/全局 | 图像 | CNN/TRANSF |
| | MCD[39] | | C-C, V | 全局 | 图像 | CNN |
| | CRAFT[40] | | C-C | 全局 | 图像 | CNN |

事后可解释方法为预训练模型的理解提供了一种低侵入性、可迁移性强的解决方案, 尤其适用于对预测精度与泛化能力要求较高的应用场景. 在不牺牲模型性能的前提下, 这类方法能够显著增强模型的透明度与用户信任度, 通过人类可理解的概念语言, 提供了一种更高层级的语义解释. 尽管如此, 事后方法也存在一定局限性. 由于概念是在模型训练完成后才被引入, 其解释行为未必反映模型在训练过程中真实采用的语义机制, 因此可能存在“贴标签式解释”的风险. 此外, 这类方法应被视为增强人类理解的补充工具, 而非对模型内部机制的完整还原. 在解释可用性与建模真实性之间取得平衡, 仍是该领域面临的重要挑战之一.

### 2.1 基于有监督概念的事后可解释方法

基于有监督概念的事后可解释方法通过利用已标注的符号化概念, 对模型进行语义层面的深入分析. 该类方法依赖一组明确定义的概念标签, 用于探究模型的预测行为是否与这些高层语义概念存在关联, 从而揭示模型的决策依据与内部机制. 通过构建模型表示与人类认知概念之间的对应关系, 此类方法有助于增强模型行为的可理解性与透明度, 提升用户对模型预测的信任度. 该类方法通常可进一步细分为两个子方向: 一类着眼于概念与类别之间的关系建模, 即分析特定概念在模型分类决策中扮演的角色; 另一类则侧重于概念与神经元激活之间的对应关系, 用于揭示模型内部表示中的语义分布结构.

在前者方向中, 概念激活向量测试方法[23]是具有代表性的工作, 尤其适用于处理有监督概念的建模任务. 如图 3 所示, TCAV 提出了一种无需修改模型结构即可评估模型是否依赖特定概念的解释框架, 具体包括以下核心步骤.

首先, 针对用户定义的每一个概念, 准备两个图像样本集: 一个为正样本集, 包含具有该概念的图像; 另一个为

负样本集, 用于表示不包含该概念的图像. 这些图像被输入待解释模型, 从中提取高层特征表示. 随后, 在特征空间中训练一个线性分类器, 用于区分正负样本, 得到该分类边界的法向量, 即概念激活向量, 用于刻画该概念在特征空间中的方向性表示. 接下来, 为评估模型预测对该概念的依赖程度, TCAV 计算模型在给定输入下的类别预测方向与 CAV 之间的点积, 即类别方向导数与概念向量之间的敏感性指标. 概念敏感度可以形式化如下:

$$S_{C,k,l}(x) = \nabla h_k(f_l(x)) \cdot v_C^l,$$

其中, $f_l(x)$ 为第 $l$ 层的激活, $v_C^l$ 为该层概念 $C$ 的概念激活向量. 该指标反映了概念在模型预测该类别时的正向影响程度. 进一步地, TCAV 定义了一个评分指标 (TCAV score):

$$TCAV_{C,k,l} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|}.$$

该值用于衡量在所有输入中, 有多少比例的样本对该概念表现出显著正向敏感性得分越高, 说明该概念对模型分类结果的影响越大, 解释效力越强. TCAV 方法的优势在于无需修改原始模型结构, 具备较强的通用性与适应性, 同时能够提供定量化的语义解释, 是目前概念-类别关系分析中的代表性方法之一. 经典的 TCAV 方法通过训练线性分类器获得概念方向 (CAV), 并利用方向导数度量概念对模型预测的影响, 为可解释性研究提供了灵活的全局分析框架. 然而, TCAV 基于向量表示, 忽略了卷积神经网络中特征图所蕴含的空间结构信息, 限制了对复杂语义模式的精细刻画. 为此, 后续本团队提出了语义激活张量方法 TSAT[41], 通过支持张量机学习概念张量, 保留通道和空间维度的结构信息, 提升了解释的语义精度与稳定性. 该方法不仅在保留模型原有结构的前提下实现了高质量的概念建模, 还显著优于 TCAV 在复杂语义概念和高分辨率图像上的可解释性能.

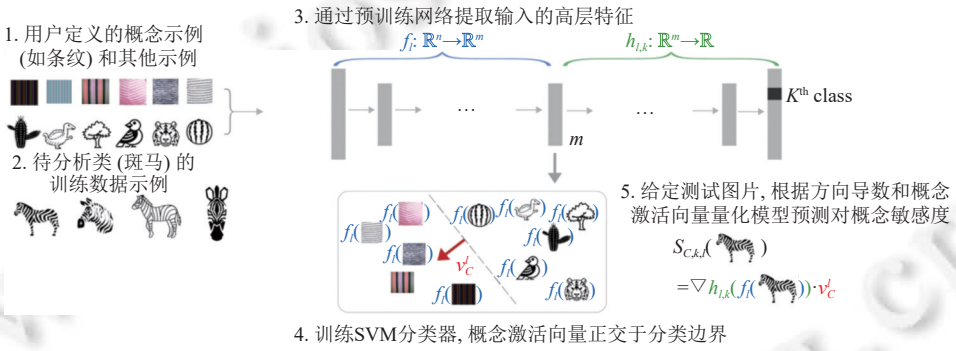


图 3 概念激活向量算法流程[23]

除了 TCAV 等方法关注概念与类别之间的相关性建模, 近年来也出现了一些更具结构性和因果性解释能力的扩展方法. 其中, IBD (interpretable basis decomposition)[30]提出了一种将预测证据分解为最相关语义组件的策略, 用于构建类-概念关系. IBD 将类别与概念均视为网络倒数第 2 层潜在空间中的向量, 并通过在该潜在空间中构建一组线性分类器来表示概念方向. 随后, 模型通过优化一个回归过程, 利用这些概念向量与样本嵌入之间的内积预测类别输出, 回归系数用于衡量各概念对类别预测的相关性强度, 而残差项则反映模型中未被概念解释所覆盖的预测变异性. 此外, IBD 还为每个概念生成对应的热力图, 以显示哪些输入区域最能激活该概念, 从而增强解释的可视化效果.

与此类“相关性驱动”方法不同, 另一类方法更进一步, 专注于分析特定概念对模型输出的因果影响. 典型代表是因果概念效应 (causal concept effect, CaCE)[31], 其核心思想是评估概念在预测中是否具有因果作用. 具体做法是对原始样本进行概念级别的干预, 生成去除某个概念后的反事实样本, 进而衡量该概念的有无对模型输出的影响差异. 该差异即被定义为 CaCE 值, 用于量化该概念的因果效应. 为了高效实现这一过程, 研究者提出了两种策略: 一是在可控环境中对样本进行人工干预, 显式添加或删除某一概念; 二是使用生成模型 (如变分自编码器) 对带或不带目标概念的图像进行建模, 从而生成近似的反事实版本. 这些策略被应用于多个数据集, 并与非因果版本的 TCAV 与基础版 CaCE 进行比较, 实验结果表明, CaCE 更有效地识别出真正与预测决策存在因果关系的概念.

除了建立类别与符号概念之间关系的方法外, 另一类研究聚焦于探索符号概念与神经网络内部神经元活动之间的映射关系, 旨在为神经元提供语义层面的解释, 揭示模型内部表征的可解释性基础. 在这一方向上, Zhou 等人[33]的研究分析了深度卷积神经网络在执行场景分类任务过程中是否能够自动学习出具备语义意义的对象检测器. 研究发现, 在训练过程中, 隐藏层神经元不仅执行特征提取任务, 还能够自主形成对特定对象或其组成部分的敏感响应. 为了更精确地量化每个神经元的语义含义, 研究者基于输入激活响应, 分析其所关联的图像区域, 并结合人类众包标注, 对神经元激活进行语义匹配与精度评估. 实验结果表明, 大多数神经元都能被映射至可识别的语义概念, 在所有隐藏层中平均精度超过 70%. 更为关键的是, 随着网络层级的加深, 所学习的概念类型从低级的颜色、纹理等逐渐过渡到更高级的对象及其组成结构, 展现出从感知到认知的语义抽象能力. 与以上工作类似, ND (network dissection)[34]系统性地提出了一种用于解释神经元语义的标准化方法. 其核心目标是将神经网络中的每个神经元激活与具体可解释概念建立明确联系. 为此, 研究者构建了一个广泛且密集标注的数据集 BRODEN, 该数据集整合了多种来源, 涵盖颜色、纹理、对象、场景等多种层级的语义标签. ND 方法通过区域重叠率 (inter-section over union, *IoU*) 来度量神经元激活图与标注概念区域之间的匹配度, 从而为每个神经元赋予最具代表性的语义标签. ND 的输出结果展示了每个神经元所响应的概念类型, 并允许多个神经元共同表征同一语义概念, 从而提升了语义表达的多样性与丰富性. 然而, 该方法的准确性高度依赖于概念标签集的覆盖范围与标注质量. 如果某些实际存在于激活图中的概念未被数据集标注, 则该神经元可能因缺乏有效对应而被视为“不可解释”, 进而影响整体解释率. 这类研究结果表明, 深度神经网络中的单个神经元在未经显式监督的情况下, 仍具有自动捕捉高层语义特征的能力. 这不仅打破了神经网络完全“黑盒”的刻板印象, 也为构建基于概念的可解释机制提供了理论依据和技术路径, 证明了神经元级别的语义对齐是可行且具有实用价值的.

### 2.2 基于无监督概念的事后可解释方法

第 2.1 节介绍的有监督概念建模方法依赖于明确的概念标签. 然而, 针对深度神经网络各层语义进行细粒度标注是一项成本高昂且极其耗时的任务, 难以适应实际场景中多变的应用需求. 在现实应用中, 为每一个新任务构建一套完整的概念标注体系往往不可行, 因此, 有必要探索无需人工监督的概念建模路径. 基于此动机, 研究者们提出了利用聚类算法对神经网络隐空间中的相似模式进行分析, 并进一步研究这些模式与模型预测结果之间的关系. 这种方法不仅能够揭示模型内部所学习到的结构性语义表示, 还在一定程度上减轻了对人工标注的依赖, 提升了可解释性方法的通用性与可扩展性.

与需要预设概念标签的 TCAV 不同, 自动概念提取方法 (automatic concept-based explanation, ACE)[36]是一种无需人工监督即可分析训练完成的分类模型中某一类别预测逻辑的技术路径. 该方法的核心流程包括以下几个阶段: 首先, 对属于目标类别的图像进行多分辨率图像分割, 以获取涵盖纹理、局部结构以及完整对象在内的多层次视觉区域. 这一过程旨在全面捕捉潜在的概念候选. 然后, 将这些图像区域嵌入到模型的高层潜在空间中, 并使用 K-means 聚类对它们进行分组, 以提取相似模式簇作为候选概念. 为避免噪声干扰, ACE 会排除那些可能导致异常预测的图像区域, 例如边缘无关区域或缺乏语义代表性的片段. 接着, 将每一个聚类得到的概念候选输入 TCAV 框架中, 计算其对特定类别预测结果的平均概念敏感度得分. 得分越高, 表示该概念对该类别预测具有更强的正向贡献. 最终, ACE 输出一个基于 TCAV 得分排序的高相关概念列表, 从而实现无监督的语义解释流程. 如图 4 所示, 该方法通过分割、嵌入、聚类和排序这 4 个阶段, 完成了从数据驱动模式提取到语义解释的完整闭环.

尽管由于图像分割、聚类或相似性度量的局限性, ACE 可能会识别出部分无意义或不一致的模式, 但通过大量人类评估实验验证发现, ACE 所提取的概念在跨样本之间具有高度一致性, 且通常可被人类观察者使用相同的自然语言标签描述. 这一发现表明, ACE 在不依赖人工标注的情况下, 仍能够发现具备认知合理性的概念结构, 成为无监督概念建模方向中的代表性方法之一.

尽管自动概念提取方法 (ACE) 能够在无需标签监督的条件下, 从神经网络中识别出概念性模式, 但其提取的概念集合未必能全面解释模型的预测行为. 为克服这一局限, Yeh 等人[37]提出了一种改进方法, 引入了“概念完整性得分”这一指标, 用于衡量一组概念对模型预测的覆盖能力. 该得分的核心公式如下:

$$\eta_f(c_1,\ldots,c_m)=\frac{Accuracy\ using\ g(vc(x))-baseline\ accuracy}{Accuracy\ of\ f(x)-baseline\ accuracy},$$

其中, $f(x)$ 表示原始模型对输入 $x$ 的预测, $vc(x)$ 表示输入在特征空间的表示, $g$ 表示将概念表示映射回输出的解释模型, *baseline* 通常指随机猜测或多数类的预测. 高完整性得分即 $\eta_f$ 越接近 1 表示该组概念能够较充分地解释模型的决策依据. 然而, 研究者进一步指出, 仅关注完整性可能不足以构建可靠的概念解释, 因为这忽略了概念的可解释性、语义一致性以及与人类直觉的契合度等因素. 为此, 作者提出了一种优化的概念发现算法, 在保证解释完整性的同时, 增强概念在潜在空间中与邻域区域的相关性, 从而提取出更具语义意义的解释单元. 此外, 借鉴了夏普利值 (Shapley value) 的思想, 提出了用于概念层的重要性评估指标——ConceptSHAP, 用于更精确地量化各个概念对类别预测的边际贡献. ConceptSHAP 分数表示如下:

$$s_i(\eta)=\sum_{S\subseteq CS\{c_i\}}\frac{(m-|S|-1)!|S|!}{m!}[\eta(S\cup\{c_i\})-\eta(S)],$$

其中, $CS=\{c_1,\ldots,c_m\}$ 代表全部概念集, 包含 $m$ 个概念; $\eta(S)$ 表示仅使用概念子集 $S$ 的完整性得分, $\eta(S\cup\{c_i\})$ 表示使用 $S$ 加上 $c_i$ 的完整性得分. 通过在合成数据集上的实证验证, 该方法在发现完备且语义合理的概念集合方面, 相较于 ACE 展现出显著优势, 为提升模型解释性与透明度提供了新的路径.

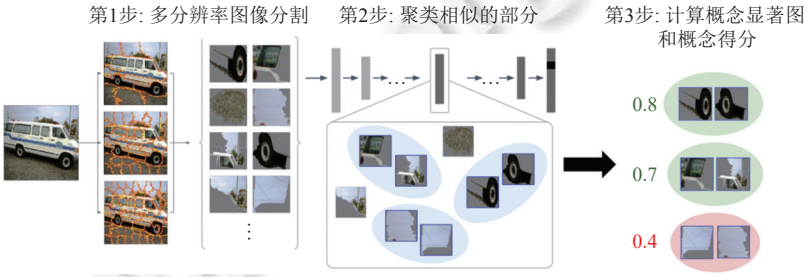


图 4 自动概念提取算法 (ACE)[36]

在 ACE 框架的基础上, ICE (invertible concept-based explanations)[38]进一步提升了概念识别与重要性评估的准确性. 该方法用非负矩阵分解替代了 K-means 聚类, 直接在特征图上提取稀疏而具有可加性的概念向量, 从而更有效地捕捉模型在分类器顶层的概念组合. 与依赖 TCAV 得分不同, ICE 将概念的重要性视为模型输出的线性近似权重, 提升了解释与模型行为的一致性. 实验表明, 与 PCA 和 K-means 等传统方法相比, ICE 在重构误差和人类可解释性评价上均表现更优.

在此基础上, MCD (multidimensional concept discovery)[39]提出通过识别深度网络隐藏空间中的多维线性子空间 (即概念子空间) 来实现全局层面的模型解释. 该方法首先使用稀疏子空间聚类 (sparse subspace clustering, SSC) 发现潜在概念, 再通过 PCA 提取每个概念簇的主方向, 并构建一个与输出层连接的线性映射, 用于计算完整性得分, 从而量化概念对预测的整体解释能力. 此外, CRAFT (concept recursive activation factorization for explainability)[40]进一步扩展了 ACE 和 ICE 提出的无监督概念挖掘框架. CRAFT 首先通过随机裁剪方式生成图像片段, 记录这些子图在模型中间层的激活响应, 并使用 NMF 进行概念因子分解, 获得高质量的概念表示. CRAFT 的一大特色是强调概念的层级结构, 采用递归方法提取概念与子概念, 并结合显著性图与敏感性分析进行可视化与重要性度量, 提升了模型多层语义结构的解释能力.

综上所述, 基于概念的事后可解释方法因其无需更改模型架构, 且能在保持预测性能的前提下提供语义层面的解释, 尤其适用于模型结构不可更动或需复用预训练模型的场景. 这些方法为研究者和工程实践提供了强有力的工具, 有助于揭示模型的内部机制、诊断预测逻辑, 并增强用户信任. 然而, 这类方法也面临一些挑战与局限. 首先, 由于其适用于已训练完毕的模型, 通常无法通过结构干预进一步提升解释性, 易受到原始架构设计的约束. 其次, 所提取的“概念”并非总能与人类语义一致, 可能导致解释的模糊性或主观性. 最后, 事后方法在鲁棒性方面普遍较弱, 容易受到对抗样本干扰, 在面对攻击时可能给出误导性解释.

基于这些问题, 越来越多研究者开始将焦点转向事中可解释方法. Rudin 等人[42]提出了一个颇具影响力的观

点: 与其在模型训练完成后费力地“解释”黑盒模型, 不如在模型设计之初就构建具备可解释性的结构, 从源头上确保模型的透明性与可理解性. 这类方法在模型设计阶段就将概念嵌入与结构可解释性作为建模目标, 使解释性与预测性在训练过程中共同优化, 从根本上提升模型的透明度、稳健性与语义一致性.

## 3 事中可解释的概念建模方法

事中可解释方法 (intrinsically interpretable method) 指的是在训练过程中引导神经网络主动学习与语义概念对齐的内部表示, 从而使网络结构本身具备内生的可解释性. 根据概念生成方式及数据依赖类型, 这类方法通常分为 4 类: ① 有监督方法依赖于明确的概念标签提供训练指导; ② 无监督方法通过网络内部结构自动形成概念表示; ③ 混合方法在此基础上融合少量标注与无监督机制, 以平衡准确性与适用性; ④ 生成方法借助外部模型构建概念嵌入, 引导模型学习更具语义深度的特征. 本文在表 2 中构建了一个多维度的分析框架, 系统整理了事中可解释方法在概念类型、解释机制、适用范围、数据类型及任务网络结构等方面的差异, 涵盖卷积神经网络、变换器网络、自编码器 (AE)[43]、变分自编码器 (VAE)[44]以及大型语言模型 (LLM) 等主流架构.

表 2 基于概念的事中可解释性方法

| 概念标注 | 方法 | 概念类型 | 解释类型 | 解释范围 | 数据类型 | 网络类型 |
|---|---|---|---|---|---|---|
| 有监督 | CBM[45] | 符号化概念 | N-C, C-C | 局部/全局 | 图像 | CNN |
| | ProbCBM[46] | | N-C, C-C | 局部/全局 | 图像 | CNN |
| | CW[47] | | N-C | 全局 | 图像 | CNN |
| | CEM[48] | | N-C, C-C | 全局 | 图像 | CNN |
| | PCBM[49] | | N-C, C-C | 局部/全局 | 图像 | CNN |
| | CT[50] | | N-C, C-C | 局部/全局 | 图像 | TRANSF |
| 无监督 | Interpretable CNN[51] | 相似模式簇 | N-C, V | 全局 | 图像 | CNN |
| | SENN[52] | | N-C, C-C | 局部 | 图像 | CNN+AE |
| | BotCL[53] | | N-C, C-C | 局部 | 图像 | CNN+AE |
| | SelfExplain[54] | | C-C, V | 局部/全局 | 文本 | CNN+AE |
| | ProtoPNet[24] | 原型 | N-C, C-C, V | 局部/全局 | 图像 | CNN |
| | ProtoTree[55] | | N-C, C-C, V | 局部/全局 | 图像 | CNN |
| | HPNet[56] | | N-C, C-C, V | 局部/全局 | 图像 | CNN |
| | TesNet[57,58] | | N-C, C-C, V | 局部/全局 | 图像 | CNN/TRANSF |
| | ProtoPShare[59] | | N-C, C-C, V | 局部/全局 | 图像 | CNN |
| | ProtoPFormer[60] | | N-C, C-C, V | 局部/全局 | 图像 | TRANSF |
| | Deformable ProtoPNet[61] | | N-C, C-C, V | 局部/全局 | 图像 | CNN |
| | MCPNet[62] | | N-C, C-C, V | 局部/全局 | 图像 | CNN |
| 混合 | CBM-AUC[63] | 相似模式簇<br>符号化概念 | N-C, C-C | 全局 | 图像、视频 | CNN |
| | Ante-hoc[64] | | N-C | 局部/全局 | 图像 | CNN+AE |
| | GlanceNets[65] | | N-C, C-C | 全局 | 图像 | CNN+VAE |
| 生成 | LaBo[66] | 文本概念 | C-C | 局部/全局 | 图像 | CNN+LLM |
| | Label-free CBM[67] | | C-C | 局部/全局 | 图像 | CNN+LLM |
| | Align2Concept[68] | | N-C, C-C, V | 局部/全局 | 图像 | CNN+LLM |

### 3.1 基于有监督概念的事中可解释方法

与事后方法类似, 事中有监督的概念建模方法也依赖于标注有符号化概念的数据集 (如每个瓶颈层神经元代表一个语义属性), 以显式引导模型中间层学习对人类可理解概念的表达. 典型代表是概念瓶颈模型 (concept bottleneck model, CBM)[44], 如图 5 所示, 该模型在输入与预测目标之间引入一个由语义概念神经元构成的“瓶颈

层”, 通过这一结构将学习过程划分为两阶段: 首先预测预定义的概念表示, 再基于这些概念生成最终输出. 该过程的数学表达式如下:

$$\begin{cases}\hat{c}=f_c(x),\hat{y}=f_y(\hat{c})\\ \hat{y}=f_y\circ f_c(x)=f_y(f_c(x))\end{cases},$$

其中, 第 1 行公式为概念预测函数表示从输入到概念的映射函数 (通常为神经网络); 第 2 行公式表示从输入到概念的映射函数 (通常为神经网络), 整个 CBM 可以表达为一个复合函数. 与传统端到端模型直接从原始输入 (如图像像素) 到预测结果 (如疾病严重程度) 不同, CBM 支持在概念层级进行解释、干预与控制. 训练阶段通过施加中间监督损失, 使瓶颈层的神经元与人类标注的概念对齐; 测试阶段则支持用户直接编辑模型预测出的概念值并传递影响至最终输出. 这一机制极大增强了人机交互能力, 允许专家如放射科医师直接修正模型对某结构的错误理解 (例如, 将正常骨组织识别为骨刺), 以此改善整体预测.

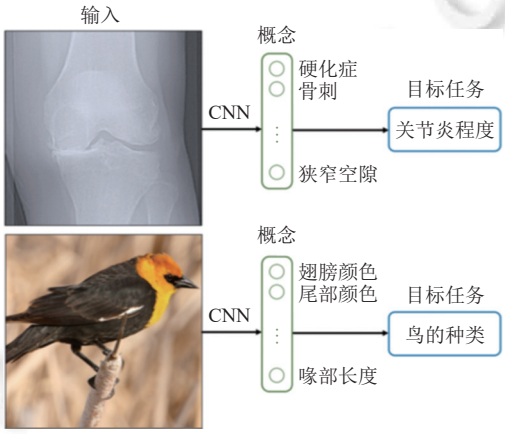


图 5 概念瓶颈层模型结构 (CBM)[44]

尽管 CBM 实现了面向概念的决策路径, 但其“确定性概念预测”在面对数据模糊性或语义歧义时可能削弱解释的稳定性与可信度. 为缓解这一问题, Kim 等人[46]提出了概率概念瓶颈模型 (probabilistic concept bottleneck model, ProbCBM), 通过对每个概念学习一个正态分布嵌入, 引入预测不确定性建模. 传统 CBM 使用确定性模型预测概念 $\hat{C}=g(x)$, 而 ProbCBM 使用概率嵌入表达每个概念:

$$p(z_c|x)\sim\mathcal{N}(u_c,diag(\sigma_c)),$$

其中, $u_c$ 和 $\sigma_c$ 由神经网络预测, 表示每个概念的均值和不确定性. 具体而言, 该模型利用蒙特卡洛采样从概念分布中估算其存在概率, 并通过这些概念嵌入进一步推断最终类别. 概念是否存在的概率计算如下:

$$p(c=1|x)\approx\frac{1}{N_s}\sum_{n=1}^{N_s}s\left(a\left(\|z_c^{(n)}-z_c^-\|^2-\|z_c^{(n)}-z_c^+\|^2\right)\right),$$

其中, $z_c^+$ 和 $z_c^-$ 分别是概念存在与不存在的锚点, $s(\cdot)$ 代表 Sigmoid 函数, $a$ 表示可学习的缩放函数. 在训练过程中, ProbCBM 联合使用二元交叉熵与 KL 散度作为损失函数; 推理阶段则可通过采样或均值估计实现. 相比原始 CBM, ProbCBM 提供了一种更加稳健、可信的语义解释方式.

然而, 传统概念瓶颈模型, 在提升解释性的同时, 常牺牲预测精度. 为解决这一权衡问题, Zarlenga 等人[48]提出了概念嵌入模型 (concept embedding model, CEM), 为每个概念学习两个具备语义意义的嵌入向量, 分别代表其激活与非激活状态. 通过对这两种向量的混合, CEM 构建出更丰富的概念表示, 并借助标签预测器完成下游任务预测. CEM 同样支持测试阶段的概念干预, 使领域专家能纠正模型对特定概念的误判, 从而调整最终输出. 此外, 作者还提出了两个评价指标: 概念对齐得分用于衡量学得的概念表示与真实标签的一致性, 信息瓶颈分析则用于评估概念表示在压缩与表达之间的权衡. 实验证明, CEM 在多个数据集 (包括 CUB 与 CelebA) 上均优于现有 CBM,

在准确性、可解释性与交互能力方面表现突出.

针对 CBM 依赖大量概念标注和性能受限的问题, Yuksekgonul 等人[49]提出了事后概念瓶颈模型 (post-hoc concept bottleneck model, PCBM), 该方法可将任意预训练模型转化为具备可解释性的 CBM 架构, 无需对训练数据进行额外标注, 从而提升灵活性和实用性. PCBM 可从外部概念数据集迁移概念信息, 或结合多模态模型从自然语言描述中提取语义表示. 其构建流程包括: 首先通过概念激活向量学习概念子空间; 然后借助文本编码器将语言描述映射至共享嵌入空间; 最后采用可解释预测器 (如稀疏线性模型或决策树) 将概念映射至最终预测. 实验结果表明, PCBM 在多个数据集上与原始模型性能相当, 而其高效变体 PCBM-h 更是在所有测试任务中匹配甚至超越原模型, 充分验证了其在无需标注条件下的可行性与有效性.

不同于以上概念瓶颈结构, Chen 等人[46]提出了概念白化 (concept whitening, CW) 技术, 通过对网络中特定层的改造, 实现潜在空间与预定义概念的对齐, 从而更清晰地揭示模型是如何逐步掌握语义概念的. CW 模块嵌入在神经网络内部, 通过白化和旋转操作对特征进行规范化, 使潜在空间的轴与人类关注的概念方向一致. 白化是一种将协方差矩阵转化为单位矩阵的线性变换, 降低特征间冗余性, 而旋转矩阵进一步将语义概念与特征轴对齐. 在训练过程中, CW 同时优化主任务损失 (如交叉熵) 与概念对齐损失, 后者通过最大化概念方向上的激活来实现对齐目标. 这种机制不仅保持了原始任务性能, 还显著增强了模型对概念的表征能力与解释效果.

以上方法中概念监督数据不必局限于主训练集本身, 也可来自外部数据源, 包括独立的概念数据集 (如 CUB-CW、CUB-CT)、半监督标注集 (如 CEM), 或结构化知识图谱 (如 PCBM). 这种扩展方法一方面执行标准的端到端训练以保证模型性能, 另一方面在特定层引入语义约束, 使模型结构中自然嵌入可解释概念, 从而构建出兼顾性能与可解释性的模型框架.

### 3.2 基于无监督概念的事中可解释方法

无监督概念模型赋予神经网络自主学习语义结构的能力, 而无需将概念与预定义符号显式对应. 通过识别特征空间中的相似模式或原型, 这类模型能够自动发现数据中的潜在结构与语义规律, 从而在不依赖人工标注的前提下提升模型的可解释性.

一类代表性方法通过对相似特征模式进行聚类, 以实现对模型特征空间的解耦. Zhang 等人[51]提出了一种方法, 将传统卷积神经网络 (CNN) 转化为结构上更具可解释性的网络, 核心在于揭示高层卷积核所学习的物体部分表示. 在该模型中, 每个高层卷积核被设计为专门响应某一特定物体区域, 且无需额外的标注信息. 作者预定义一组仅响应局部区域的模板, 并以模板与卷积核之间的互信息作为损失函数, 鼓励每个卷积核学习并编码特定语义区域, 避免不同卷积核对相同区域的冗余激活. 实验结果表明, 该方法在多个 CNN 架构和数据集上均显著提升了物体部分的定位稳定性与可解释性.

另一个方向是通过模型架构设计实现自解释能力. Alvarez-Melis 等人[52]提出了自解释神经网络 (self-explaining neural network, SENN), 该框架从线性模型出发扩展至更复杂的深度结构. SENN 通过一个概念编码器提取输入特征并进行聚类; 然后使用参数化模块根据输入生成每个概念的相关性权重; 最终以概念与其权重的线性组合输出预测结果. 为提升表达稀疏性与概念可辨性, 作者引入稀疏正则化, 确保每个输入仅依赖少量不重叠的概念. SENN 实验结果表明, 该模型在保持分类性能的同时, 能提供结构清晰、解释直观的概念表达. 不过, SENN 的设计更偏向于学习覆盖全局的概念表达, 缺乏对局部解释的支持. 为此, bottleneck concept learner (BotCL)[53]进一步引入槽注意力机制 (slot attention), 以自动识别图像中的局部概念表示. BotCL 在重建损失的基础上引入对比损失, 增强跨类概念的辨别能力, 促进语义聚类的明确性. 实验证明, 在多个图像分类任务中, BotCL 在准确性和解释性上均优于包括 SENN 在内的多种基线模型. 该模型不仅能够自动识别关键局部概念, 还能通过槽注意力机制对其进行可视化, 例如在 ImageNet 上识别鲨鱼的口部和鳍部, 从而构建结构化的图像语义表示.

在自然语言处理领域, 研究者也在探索可解释模型的构建. Rajagopal 等人[54]提出的 SelfExplain 模型, 将全局与局部可解释机制融合进文本分类器中, 以短语级概念替代传统词级特征作为解释单位. 该模型一方面通过最大内积搜索检索训练集中对当前输入最具代表性的全局概念, 另一方面评估这些局部概念与目标标签之间的相关性, 从而量化其重要性. 在 5 个主流文本分类数据集上, SelfExplain 展现出在不损失性能的情况下, 显著增强解释

效果的能力. 与现有基线相比, 其生成的解释不仅合理可信, 更在人工评估中表现出较高的一致性与可理解度.

在无监督概念可解释方法中, 另一重要方向是基于原型的概念建模. 该类方法通过编码训练样本的局部区域, 自动学习具备代表性的原型概念, 并基于其与输入样本的相似度完成分类. 与显式的符号标签不同, 原型学习方法赋予模型从特征空间中抽取“具象例子”的能力, 不同研究对原型施加了多样的结构约束与语义属性, 从而呈现出更丰富的可解释机制.

Chen 等人[24]在 2019 年提出的原型部件网络 (ProtoPNet) 是该方向的开创性工作, 框架图如图 6 所示. 该方法通过识别图像中典型部位并聚合原型证据进行分类, 模拟了人类识别中的“比对典型特征”过程. ProtoPNet 利用卷积特征 $z \in \mathbb{R}^{H\times W\times D}$ 与原型 $p_j \in \mathbb{R}^{1\times 1\times D}$ 之间的相似度生成激活图, 原型层计算公式如下.

$$g_{p_j(z)} = \max_{\tilde{z}\in patches(z)} \log\left(\frac{\|\tilde{z}-p_j\|_2^2+1}{\|\tilde{z}-p_j\|_2^2+\varepsilon}\right).$$

该式得分越高, 代表图像中某部分越接近该原型. 原型激活图通过全局最大池化获得概念匹配得分, 最终以线性加权输出预测结果. 经典的 ProtoPNet 网络训练过程分为 3 个阶段.

● 阶段 1, 原型表示学习: 优化卷积网络 $f$ 和原型层 $\boldsymbol{P}$, 固定全连接层权重 $\omega_h$. 该阶段使用交叉熵损失、类内聚集损失和类间分离损失联合优化, 其中 $\boldsymbol{P}$ 为集合, $\boldsymbol{P}=\{p_j\}_{j=1}^m$.

$$\begin{cases}\min\limits_{\boldsymbol{P},\omega_{\text{conv}}} \dfrac{1}{n}\sum\limits_{i=1}^{1} CrossEntropy(h\circ g_p\circ f(x_i),y_i)+\lambda_1 Clst+\lambda_2 Sep\\ Clst=\dfrac{1}{n}\sum\limits_{i=1}^{n}\min\limits_{j:p_j\in P_{y_i}}\min\limits_{z\in patches(f(x_i))}\|z-p_j\|_2^2;\ Sep=-\dfrac{1}{n}\sum\limits_{i=1}^{n}\min\limits_{j:p_j\notin P_{y_i}}\min\limits_{z\in patches(f(x_i))}\|z-p_j\|_2^2\end{cases}.$$

● 阶段 2, 原型投影: 将每个原型 $p_j$ “推”到与其最近的训练样本 latent patch 上, 使其可视化:

$$\begin{cases}p_j\leftarrow \arg\min\limits_{z\in\mathcal{Z}_j}\|z-p_j\|_2\\ \mathcal{Z}_j=\{\tilde{z}:\tilde{z}\in patches(f(x_i)),\ \forall i,\ \text{s.t. } y_i=k\}\end{cases}.$$

● 阶段 3, 基于原型的分类: 使用带 $l_1$ 正则项调整全连接层权重 $\omega_h$, 鼓励基于自身类别原型进行分类:

$$\min_{\omega_h}\frac{1}{n}\sum_{i=1}^{n} CrossEntropy(h\circ g_p\circ f(x_i),y_i)+\lambda_3\sum_{k=1}^{K}\sum_{j:p_j\notin P_k}\left|\omega_h^{(k,j)}\right|.$$

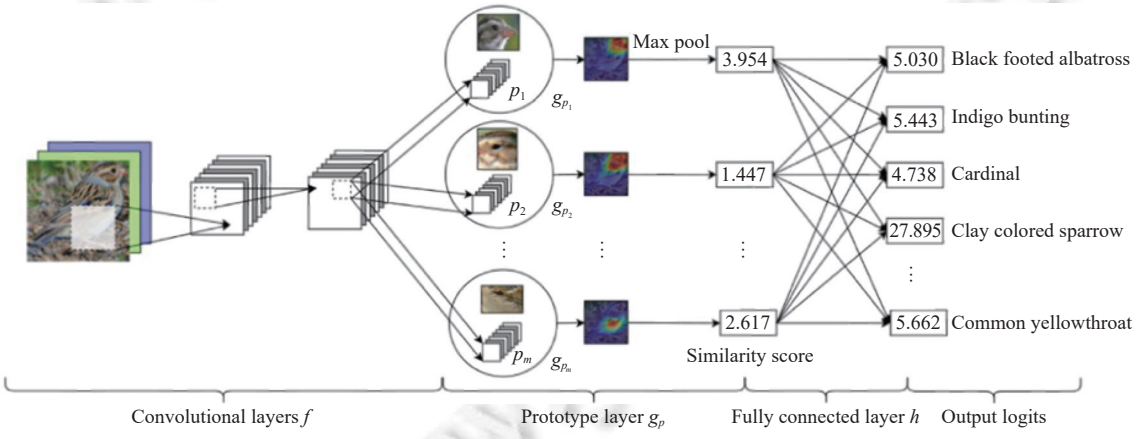


图 6 ProtoPNet 网络框架图[24]

图 7 中展示了 ProtoPNet 模型在图像分类中的推理过程, 该模型通过原型比对的方式实现可解释性分类. 具体而言, 输入图像首先经过卷积网络提取特征, 再与预先学习到的多个原型进行匹配, 每个原型代表某类典型局部特征 (如鸟的喙或羽毛纹理). 模型在输入图像中定位与各原型最相似的区域, 并计算对应的相似度分数; 随后将这些分数与类别连接权重结合, 得到每个原型对该类别的支持度. 所有原型的支持度相加形成该类别的总分, 最终由总分最高的类别作为模型输出.

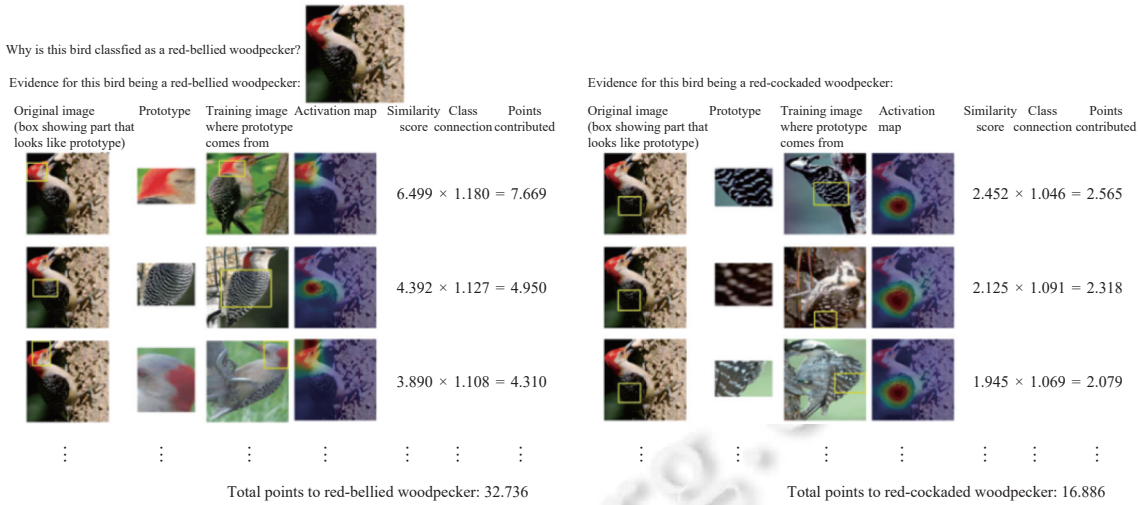


图 7 ProtoPNet 推理过程[24]

为增强模型结构的层次性, Nauta 等人[55]提出了原型树网络 (ProtoTree), 将原型建模为二叉树节点, 构建从根到叶的显式分类路径. 每个内部节点包含可学习原型, 其存在性决定图像的路径选择, 叶节点则负责学习类别分布. ProtoTree 引入剪枝技术和原型可视化机制, 有效增强了全局解释性. Hase 等人[56]进一步提出分层原型网络 (HPNet), 支持从粗到细的多级原型组织结构, 使模型不仅能判断“该图像为何属于某类别”, 还能解释其在层级分类体系中的归属关系. HPNet 能在处理未见类别 (如“手枪”) 时输出合理推断 (如归类为“武器”), 提升了模型泛化能力.

在对 ProtoPNet 框架的进一步拓展中, 本团队提出了 TesNet (transparent embedding space network)[57,58], 通过构建透明嵌入空间学习概念表示, 显著增强了原型之间的判别性与解释能力. TesNet 将每个原型视为子空间中的基向量, 不同类别的原型子空间在 Grassmann 流形 (Grassmann manifold) 上进行投影, 使得类间分布更加离散可分. 该方法还引入原型可视化机制, 将原型与图像中的关键区域进行显性对应, 提升语义解释的直观性和可信度. TesNet 在多个细粒度图像分类任务中显著提升了分类性能与解释质量. 此外, 其后续的期刊拓展版进一步扩展至 Transformer 架构, 并引入了用于缓解概念冗余的剪枝算法, 提升概念压缩率与语义紧凑性. 同样为缓解 ProtoPNet 中原型冗余与类别隔离的问题, Rymarczyk 等人[59]提出 ProtoPShare 方法, 在适配 Vision Transformer 架构时, Xue 等人[60]提出 ProtoPFormer 结合全局与局部原型建模机制, 全局原型提供对象的整体表示, 局部原型专注前景区域, 两者协同校正, 强化了模型对判别性区域的关注.

此外, 为克服原型匹配在几何变换下的脆弱性, Donnelly 等人[61]设计了可变形原型网络 (deformable ProtoPNet), 将原型表示为可自适应空间位置的部件组合, 从而增强了模型对目标尺度与姿态变化的适应能力, 在多个主干 CNN 上均获得显著性能提升. 考虑到概念的层次性, MCPNet (multi-level concept prototype network)[62]则进一步引入多层次语义原型机制, 通过在多个网络层中学习局部、中层与全局原型, 并利用概念分布损失和类别对齐约束提升了原型的可区分性与一致性. MCPNet 不依赖外部标注, 兼容主干网络架构, 并支持基于原型路径的可视化解释, 在 few-shot 场景与多数据集实验中均展现出强大解释能力与准确率.

### 3.3 基于混合概念的事中可解释方法

融合监督与无监督学习范式的混合型事中可解释方法, 在监督信息稀缺的条件下, 依然能够构建出具备语义解释能力的深度模型. 通过综合两种学习机制的优势, 这类方法提升了模型在低标注场景下的适应性和可解释性, 增强了对模型行为的理解能力.

为缓解 CBM 在概念数量有限时性能下降的问题, Sawada 等人[63]提出了 CBM-AUC, 结合了传统 CBM 与优化后的 SENN 结构, 以同时学习监督概念与无监督概念. 针对原始 SENN 编码器-解码器结构在处理大规模图像时计算效率低的问题, 作者设计了 M-SENN 架构: 首先, 用鉴别器替代解码器, 训练目标转为最大化输入数据与中间

特征的互信息, 避免传统重构误差目标对学习的不利影响; 其次, 引入中间网络共享机制, 让编码器与参数化器共用特征提取模块, 从而有效压缩模型参数并提升效率. CBM-AUC 在多个任务上展现出优于原始 CBM 和 SENN 的准确性, 且通过 Grad-CAM 可视化分析, 验证了模型能准确聚焦于语义关键区域, 无论是监督还是无监督概念.

此外, Sarkar 等人[64]提出了一种通用框架, 可将“自我解释模块”无缝嵌入任意深度神经网络流程中. 该方法核心为一个概念编码器, 负责从特征中学习用于解释模型预测的语义概念, 并配合重建损失和保真度损失共同优化, 以增强解释的信息承载力与忠实性. 在 CIFAR-10、ImageNet、AwA2 和 CUB-200 等数据集上, 该方法不仅保持了与标准分类模型相当的准确率, 同时生成了语义一致、具有洞察力的解释, 且在有监督和无监督场景下均超越多个现有基线, 展示了良好的通用性.

与前述研究聚焦于模型性能或集成方式不同, Marconato 等人[65]提出的 GlanceNets 则旨在提升概念表示的质量与鲁棒性. 该方法采用变分自编码器替代传统自编码器, 引入解耦表示学习与开放集识别机制, 强化模型对高层语义概念的建模能力. 为应对“概念泄漏”问题 (即模型学习到与任务无关的伪概念), GlanceNets 在测试阶段借助开放集识别策略, 有效筛除不符合训练分布的输入样本, 提升预测的可靠性. 此外, 该方法支持多级监督粒度, 既适用于弱监督甚至无监督环境, 又可通过显式概念监督增强语义一致性和模型可解释性.

### 3.4 基于生成概念的事中可解释方法

随着大语言模型 (LLM) 的快速发展, 其丰富的语言知识为构建概念瓶颈层提供了新的可能性. Yang 等人[66]提出的 LaBo (label bottleneck) 模型便是这一方向的代表性工作. 如图 8 所示, LaBo 利用 GPT-3 生成候选语义概念, 并通过子模函数进行筛选优化, 构建具有判别力的概念瓶颈层. 具体流程包括: 首先定义图像分类任务的数据集, 并利用预训练的多模态模型 (如 CLIP[69]) 将图像和文本映射至共享语义空间; 然后基于子模函数从 GPT-3 提供的大量候选中选出最优概念集合; 接着建立从图像特征到概念得分的映射; 最后通过线性函数将概念得分转换为最终分类结果. 在包含常规物体、细粒度物体、动作、纹理、皮肤病变和卫星图像等共计 11 个数据集上, LaBo 均展示出在小样本条件下优于传统端到端模型的性能, 并在大样本设置中仍保持良好竞争力, 展现了其在提升可解释性的同时兼顾准确性的潜力.

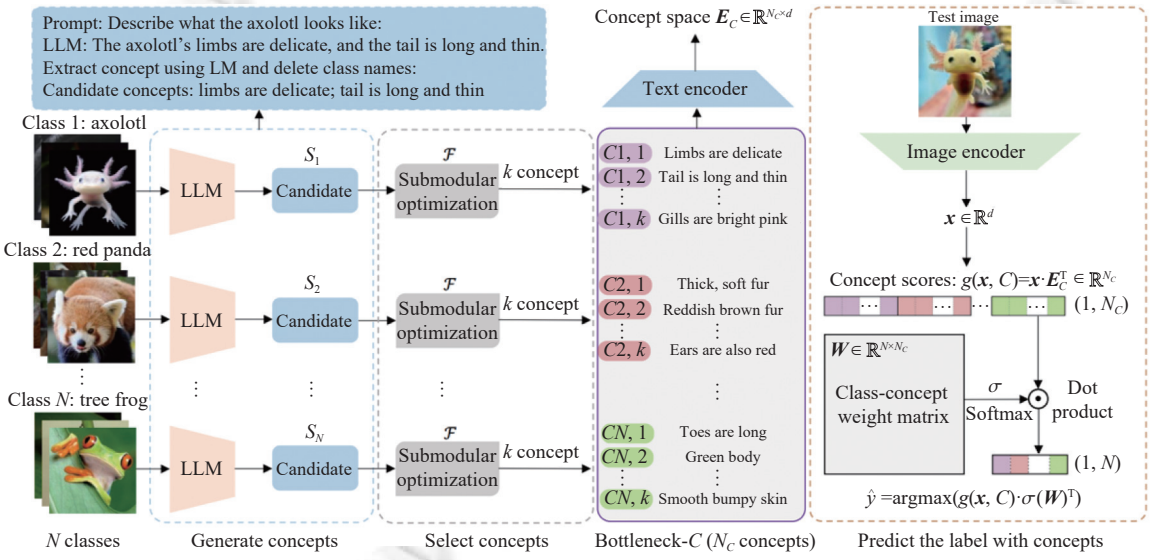


图 8 LaBo 模型结构[66]

与 LaBo 类似, Label-free CBM[67]也采用 GPT-3 自动生成初始概念集. 其方法首先基于类别名称, 通过提示生成相关描述性文本并提取潜在概念; 随后通过一系列过滤规则进行概念筛选, 如剔除冗长概念、去除与类别过于接近或相互重复的概念、筛除训练集中不常见的无效概念等. 优化后的概念集合通过 CLIP-Dissect 训练获得概念

瓶颈层的映射权重, 并通过稀疏化输出层完成最终分类. 该方法有效提升了模型可解释性, 同时保持与黑盒模型相当的性能, 进一步证明了语言模型在无监督或弱监督场景中构建概念空间的可行性.

与 LaBo 等依赖多阶段后处理的文本概念构建方法不同, 本团队提出的 ProCoNet[68] 在文本生成环节引入了基于 in-context learning 的提示策略, 借助 GPT-3 直接生成高度聚焦于视觉属性的文本概念. 该策略无需额外的拆解与过滤步骤, 有效提升了生成文本的相关性与效率. 这些自动生成的文本概念不仅能精准对应模型识别到的视觉原型, 还使模型能够同时提供“看起来像什么” (视觉层面) 与“像什么” (语言层面) 的双重解释. 在认知心理学双重编码理论的启发下, ProCoNet 被设计为一种融合视觉原型与语言概念对齐的多模态可解释图像识别框架, 构建了如下 3 个核心表示空间. 视觉空间: 用于提取图像的局部原型特征, 基于 ProtoPNet 架构获取显著区域的视觉原型; 文本概念空间: 利用 GPT-3 自动生成具有视觉可识别性的描述性短语, 并通过 CLIP 的文本编码器映射至嵌入空间; 多模态空间: 由预训练的 CLIP 模型提供跨模态对齐能力, 用作视觉与文本之间的桥梁. 由于视觉原型与文本概念嵌入分别来自图像与语言模态, 原始分布位于不同的流形 (manifolds) 上. ProCoNet 在 Stiefel 流形上优化投影矩阵, 通过 Cayley 变换将视觉原型投影至多模态空间中与文本概念对齐, 同时保持其局部几何结构的稳定性. 这一设计不仅提升了语义对齐的一致性, 还避免了对人工标注或测试阶段图像裁剪的依赖, 从而增强了模型的可解释性与泛化能力.

总体来看, 事中概念模型通过在网络结构中显式集成概念表示, 赋予模型更高层次的可解释性. 这种架构天然具备透明性强、可交互性好、语义契合度高等优势, 使得用户不仅可以理解模型的预测结果, 还能在概念层面进行干预和反事实推理, 尤其适用于医疗、金融等对合规与可控性要求较高的场景. 然而, 事中模型也存在若干挑战. 首先, 为了保证可解释性, 其结构往往受到限制, 可能牺牲一定的性能, 尤其在表达复杂模式时受限; 其次, 与事后解释方法相比, 事中方法在建模初期需设计概念结构与机制, 增加了实现难度与对专业知识的依赖; 再次, 部分模型的解释粒度有限, 难以覆盖决策全过程或应对任务迁移时的泛化需求. 因此, 尽管事中概念模型在提升语义透明性方面具有显著优势, 如何在解释性、性能与泛化能力之间实现更优平衡, 仍是未来研究的重要方向.

## 4 可解释性算法性能评估方法

为了科学地评估基于概念的可解释人工智能方法的有效性与实用性, 研究者提出了多种性能评估策略. 这些方法不仅用于衡量模型解释的质量和可信度, 也有助于理解概念与模型行为之间的关系. 总体来看, 现有的可解释性算法评估手段可划分为定性指标、定量评估.

### 4.1 定性评估指标

定性评估侧重于通过可视化与人工分析的方式, 检验模型解释是否具备语义清晰性、感知一致性与人类直观可理解性. 这类方法通常不依赖具体的数值度量, 而是强调解释的直观合理性与使用者的主观接受程度, 是概念可解释性研究中不可或缺的重要环节.

(1) 原型可视化[24,55]: 在原型网络 (如 ProtoPNet、ProtoTree) 中, 研究者通常通过激活图或原型图像 patch 可视化模型学习到的概念. 直观展示模型“看到了什么”并据此做出预测, 有助于判断这些原型是否具备一致的视觉特征. 例如, 如果多个图像中的原型部分均对应鸟类的“翅膀”, 则说明该概念具备一致性和可解释性. 通过这种方式可视化“概念对预测的贡献区域”, 有助于评估概念语义与图像内容的一致性.

(2) 人工评估[52]: 基于人类的主观认知, 强调从人类用户的视角出发判断模型生成的概念是否具备清晰的语义意义和直观的可解释性, 主要可以分为人工命名一致性、可理解性评分、语义聚类可感知性. 人工命名一致性: 由多位评审者对自动提取的概念进行命名, 统计命名的一致性指标以评估概念的语义清晰度. 可理解性评分: 邀请用户或领域专家对每个概念的可感知性、合理性和实用性打分, 反映其对人类认知的契合程度. 语义聚类可感知性: 分析不同样本被聚类为同一概念时的主观一致性, 即“是否看起来属于一类”.

### 4.2 定量评估指标

定量评估方法侧重于以数值方式衡量可解释性模型的解释质量、可信度与任务表现, 此类方法通常依赖模型

中间层输出、标注信息或辅助模块. 相较于定性评估的主观性, 这类方法具有可复现性和对比性, 主要包括以下几个关键维度.

(1) 概念准确度[45]: 在 CBM 等显式建模语义概念的模型中, 概念准确度用于衡量模型中间层对语义概念的预测质量, 是评估概念可解释性的重要基础指标. 该指标依赖具有明确概念标签的数据集, 反映模型对预定义概念的学习能力与表达精度. 具体度量方式可根据概念的类型有所不同, 对于分类型概念, 常采用 0-1 错误率或 $F1$ 分数来评估模型对离散语义属性 (如颜色、形状等) 的识别准确性; 对于连续型概念, 则通常使用均方误差 (MSE) 或均方根误差 (RMSE) 来衡量模型在概念数值预测中的偏差程度.

(2) 概念部位定位距离[58]: 该指标用于衡量模型中学习到的概念是否具备明确的视觉对应性, 能否准确定位到图像中的语义部件. ProtoPNet 等原型网络中可以广泛应用, 通过激活图中最大响应位置 $\kappa\left(b_j^{(c)}\right)$ 推理出原图中预测的 landmark 位置 $\hat{P}\left(b_j^{(c)}\right)$, 计算其与真实 landmark 之间的归一化欧氏距离:

$$\begin{cases}\hat{P}\left(b_j^{(c)}\right)=\kappa\left(b_j^{(c)}\right)\cdot\delta_{\text{jump}}+\delta_{\text{center}}\\ d_{x_i}\left(\hat{P}_k^{(c)},\hat{P}\left(b_j^{(c)}\right)\right)=\dfrac{\|\hat{P}_k^{(c)}-\hat{P}\left(b_j^{(c)}\right)\|^2\cdot}{\sqrt{\bar{w}^2+\bar{h}^2}}\end{cases}$$

(3) 概念冗余度[58]: 概念冗余度衡量所学概念之间的相似性重叠程度, 用以检验模型是否存在多个原型表达相同语义的问题. 为避免模型学习出多个语义重叠的概念, 引入冗余率指标 $\rho^{(c)}$ 来度量不同概念之间的独立性. 若多个概念被分配至同一 landmark, 则认为存在冗余. 冗余率定义如下:

$$\rho^{(c)}=1-\frac{\#unique\ predicted\ landmark}{\#prototypes},$$

其中, $\rho^{(c)}$ 数值越高, 表示概念之间的冗余越严重, 模型的语义表达多样性较差.

(4) 概念对分类预测的影响: 除了关注模型的整体分类准确率外, 评估特定概念是否真正对模型的最终预测产生影响, 也是衡量可解释性质量的重要维度. 典型的定量指标包括: TCAV 分数, 用于衡量模型对某一概念的预测敏感性; CaCE[31]通过构造反事实样本, 定量分析概念存在与否对预测结果的因果影响; ConceptSHAP[37], 该方法基于 SHAP 框架, 评估每个概念在预测输出中的边际贡献. 这些方法从不同角度揭示了模型在决策过程中是否真正“使用”了人类可理解的语义概念. 此外还有一类指标可以度量概念表示是否有助于提升模型的整体预测精度和泛化能力. 常见的评估方式包括: 概念完备得分用于衡量概念子空间在多大程度上能够解释原始模型的决策; 保真度表示基于概念的预测在准确率上是否能够逼近或重构原始模型的输出.

(5) 概念与神经元的关系: 在相关研究中, 研究者通常构建具备像素级语义标注的数据集[34,35], 以评估深度神经网络中单个神经元与具体语义概念之间的对齐关系. 常用的评估指标为 $IoU$, 用于衡量某一神经元激活区域与语义概念标注区域之间的重叠程度, 从而量化该神经元对特定概念的专注程度与表达能力. 公式表达如下:

$$IoU_{k,c}=\frac{\sum_x|M_k(x)\cap L_c(x)|}{\sum_x|M_k(x)\cup L_c(x)|},$$

其中, $M_k(x)$ 表示在图像 $x$ 上第 $k$ 个神经元单元的二值激活区域; $L_c(x)$ 表示图像 $x$ 上关于概念 $c$ 的像素级语义标注掩码.

(6) 概念解释的鲁棒性: 传统评估指标, 如概念准确度、$IoU$ 等, 只衡量解释与真值之间的静态吻合度, 却无法量化解释在扰动或多次推理中的“自我一致性”. 换言之, 这些指标回答的是“解释有多准”, 却忽略了“解释是否始终如一”. 概念解释的鲁棒性正是聚焦于解释本身 (显著图、概念权重、注意力分布等) 在面对输入扰动、模型随机性或推理路径变化时的稳定性. 其公式如下:

$$Robust(x,x')=sim(e(x),e(x')),$$

其中, 原始输入 $x$ 的解释向量 (显著图) 为 $e(x)$, 扰动输入 $x'$ 的解释向量为 $e(x')$, 通常 $sim$ 常取余弦相似度.

(7) 以人为中心的评估框架: 可解释性本质上是一项“为人而生”的特质. 元预测器 (meta-predictor)[70] 框架将人置于评估闭环的核心——通过测量用户在获得解释后, 能否准确预测模型的后续行为来直接量化解释的实用价

值, 从而规避了偏见等问题. 具体地, 通过以下公式来把“解释对人有没有用”量化成一个 *Utility-K* 值:

$$Utility\text{-}K = \frac{\mathbb{P}(\psi^{(K)}(x) = f(x))}{\mathbb{P}(\psi^{(0)}(x) = f(x))},$$

其中, 分子表示用 $K$ 个带解释的样本训练后, 人在新样本上能猜对模型结果的概率, 分母表示完全不看解释时的概率, 人猜对的概率比值 >1 就说明解释确实帮到人; =1 说明没用; <1 反而更差. 实验时 $K$ 取多个值, 画出 *Utility-K* 曲线, 再算曲线下面积 (AUC) 得到最终 Utility 分数. 作者基于 1 150 名受试者的大规模心理物理学实验, 在 3 个真实场景中系统检验了主流归因方法的人本效用, 结果一致表明: 传统的“忠实度”指标与人类的实际收益之间几乎不存在有效关联.

尽管已有多种指标被提出并广泛使用, 但整体来看当前可解释性评估体系仍处于早期阶段, 评估指标零散、定义不统一, 且大多依赖特定模型或任务. 未来应致力于构建标准化的评估基准与统一的评价体系, 以支持不同模型间更公平、可比的解释能力对比, 从而为构建可信、可控的智能系统打下坚实基础.

## 5 未来展望与挑战

尽管基于概念的可解释深度学习方法在提升模型的语义透明性和人类可理解性方面取得了诸多进展, 但该领域仍处于快速演进阶段, 面临多方面的挑战, 仍有广阔的研究空间亟待深入探索. 现有方法在概念建模、理论支撑、模态对齐、模型规模适配等方面尚存诸多不足. 为了推动可解释性研究从局部实验走向系统化发展, 以下几个方向可能成为未来突破的关键路径.

(1) 语义异构性与概念建模的复杂性. 在现实物理世界中, 任务场景千差万别, 概念空间随之呈碎片化分布. 医疗影像需要捕捉“坏死灶边缘毛糙度”这类微米级特征; 自动驾驶则要同时解析“信号灯相位-行人意图-车辆轨迹”的多层耦合语义. 二者在概念粒度、抽象层级与专业知识壁垒上差异巨大. 目前主流研究仍聚焦于简单图像分类, 依赖预定义概念即可勉强应对, 一旦转向复杂场景, 静态概念集既难以穷尽隐含模式, 又无法跟随数据演化 (如 COVID-19 影像特征持续变异), 导致解释缺口不断放大. 因此, 亟需构建自适应概念发现机制, 依托无监督或自监督学习, 从原始数据中自动蒸馏概念基元, 摆脱对人工先验的依赖. 更进一步, 可打造可生长的概念图谱——底层由数据驱动生成细粒度原型, 顶层由大语言模型实时映射为人类可读的语义标签, 形成分层锚定; 同时内置增量扩展与概念遗忘模块, 当数据分布漂移或医学指南更新时, 系统自动淘汰失效概念、吸纳新兴模式, 使概念空间始终与物理世界同步演化.

(2) 概念解释的基础理论研究亟待突破. 当前可解释人工智能中的“概念解释”研究面临理论瓶颈. 首先, 神经网络生成的连续向量空间与人类心智赖以运作的离散符号体系之间存在难以弥合的认知鸿沟——前者基于高维数值分布, 后者依赖结构化符号推理, 两者在表示机制和操作规则上存在本质差异. 其次, 现有研究缺乏对“概念”本质的深层理解, 大量工作停留在围绕 ProtoPNet、CBM 等模型的工程化改进, 却鲜少追问概念在认知层面的真正内涵. 最后, 缺乏统一的理论框架来定义何为“好的概念解释”, 导致不同方法之间难以比较, 解释质量评估缺乏客观标准. 破解这一难题需要从认知对齐的角度重新审视概念解释问题, 核心思路是构建“神经-符号”桥梁, 借鉴认知科学中的“原型理论”[71]和“概念空间理论”[72], 将人类概念建模为概率原型与语境依赖边界的统一体. 具体而言, 可构建一个 3 层映射理论框架: 神经表示层, 刻画深度网络中的分布式特征表示, 通过高维向量空间编码输入的底层特征模式; 概念抽象层, 建立连接神经表示与符号概念的中间抽象, 采用概念空间理论[72]的几何结构将概念表示为具有凸性质的区域, 其中原型对应区域中心, 边界体现概念的模糊性和语境敏感性; 认知理解层, 对应人类的概念认知结构, 确保抽象概念与人类直觉和专家知识保持一致. 在技术实现上, 通过信息论原理量化不同层间的信息传递损失和压缩效率, 运用拓扑几何学描述概念在高维空间中的邻域结构、连通性和可分性, 结合认知心理学实验验证概念分类的生态有效性, 从而确保概念表示既具备数学严谨性, 又符合人类认知规律.

(3) 多模态概念建模与对齐机制亟需完善. 多模态概念建模面临多重挑战. 首先, 模态异构性导致的表示鸿沟. 视觉信息以像素矩阵编码, 语言信息以符号序列表达, 听觉信息以频谱特征呈现, 这些异构表示在数学结构、信息密度和时空特性上存在本质差异, 难以直接建立映射关系. 其次, 语义粒度的多尺度不匹配. 视觉概念可能对应像

素级细节 (“黑色的翅膀尖端”)、区域级特征 (“黄色的喙”) 或全局语义 (“飞翔的鸟”), 而语言描述的抽象层次往往与视觉特征的空间尺度不一致, 造成概念边界模糊. 最后, 缺乏统一的对齐评估标准. 现有方法多采用任务特定的对齐策略, 缺乏跨模态概念质量的统一度量框架, 导致不同模态概念之间的一致性难以保证. 解决多模态概念对齐的核心思路是构建认知启发的统一语义空间, 可以借鉴人类多感官整合的认知机制, 采用多阶段对齐策略. 首先是模态内概念提纯, 在各模态内部通过自监督学习提取具备认知合理性的概念原型; 其次是跨模态语义锚定, 利用共享的认知概念 (如基本层次概念) 作为锚点, 建立模态间的语义对应关系; 最后统一空间投影, 将不同模态的概念表示投影到共享的几何语义空间中, 保持概念的拓扑结构和语义关系. 此外可以引入人机交互方法, 实时展示概念在不同模态的邻域, 让人类用点击拖拽微调概念边界, 每一次交互被记录为软约束, 使不同模态的概念表示在人类认知的监督下进行准确对齐.

(4) 大模型时代下的可解释性研究难题. 大模型可解释性面临前所未有的挑战. 首先, 参数规模与结构复杂性的指数级增长. 从 GPT-3 的 1 750 亿参数到 GPT-4 的万亿级参数, 模型内部包含许多注意力头和前馈层, 其决策路径呈现高度非线性和多层级嵌套特征, 传统的神经元级分析方法在计算复杂度和解释粒度上都面临瓶颈. 其次, 黑盒访问限制下的信息不对称. 商业化大模型通常只提供 API 接口, 研究者无法获取中间层激活、梯度信息或注意力权重, 这使得基于内部状态分析的传统解释方法失效. 最后, 生成式特征导致的解释验证困难. 大模型的输出具有高度随机性和创造性, 同一输入可能产生截然不同的输出, 这使得解释的一致性、可重现性和因果有效性难以保证. 针对大模型黑盒特性, 可解释方法可以从重构模型的概念推理过程和理解模型内部角度着手. 研究人员尝试将表层概念显化, 如利用思维链 (CoT)[73]和提示词工程[74] 将隐式推理路径转化为“文本化的概念序列”, 使中间推理步骤变得可观察和可验证. 此外还通过对比性提示、反事实干预和概念激活探测等方法[75], 系统性地探索模型的概念知识边界和推理模式. 还有一种研究思路利用更强的语言模型 (如 GPT-4) 对目标模型进行“概念标注”, 构建模型内部的概念知识图谱. Meta 最新提出的 large concept model (LCM)[76] 把“概念”明确定义为句子级语义单元, 将传统的“下一个 token”预测升级为“下一个句子”预测. 通过在句子嵌入空间内完成全部推理, LCM 天然获得一组可读、可写的高层语义变量, 无需 CoT 提示就能直接输出概念层面的解释, 从而有效抑制链式思维中常见的幻觉. 综上, 大模型可解释性研究仍处于萌芽期, 传统解释范式在超大规模参数、黑盒访问限制与生成不确定性等问题面前全面失效, 亟需建构适配大模型的新型解释范式.

## 6 结束语

随着深度学习在关键任务中日益广泛的应用, 其“黑盒”特性所带来的安全性、可控性与信任危机日益凸显. 基于概念的可解释方法, 作为连接模型表示与人类认知的重要桥梁, 正成为当前可解释人工智能研究的热点与前沿方向. 本文围绕“概念建模”这一核心视角, 系统梳理了深度模型中概念建构的主要形式、建模机制与应用路径, 按照事后与事中两个范式以及对概念标注的需求程度对代表性工作进行了归类与比较, 揭示了不同方法在语义表达、推理透明性与认知一致性方面的优势与局限.

总体而言, 概念不仅提升了解释的语义深度, 更推动了从“模型关注了什么”到“模型理解了什么”的转变. 未来, 随着认知科学、大模型技术与因果建模方法的持续融合, 基于概念的可解释研究有望迈向更高层次的语义抽象、更强的因果推理能力与更优的人机协同性能. 期待这一研究方向在理论深化与实际落地中持续突破, 助力构建更加透明、可信与智能的深度学习系统.

## References

[1] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

[3] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[4] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2021.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[6] Zhang S, Yao LN, Sun AX, Tay Y. Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 2019, 52(1): 5. [doi: 10.1145/3285029]

[7] Chib PS, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: A survey. IEEE Trans. on Intelligent Vehicles, 2024, 9(1): 103–118. [doi: 10.1109/TIV.2023.3318070]

[8] Teng QY, Liu Z, Song YQ, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems, 2022, 28(6): 2335–2355. [doi: 10.1007/s00530-022-00960-4]

[9] Biswas AA. A comprehensive review of explainable AI for disease diagnosis. Array, 2024, 22: 100345. [doi: 10.1016/j.array.2024.100345]

[10] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 2020, 3: 26. [doi: 10.3389/frai.2020.00026]

[11] Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin SF. Explainable AI in credit risk management. arXiv:2103.00949, 2021.

[12] Omeiza D, Webb H, Jirotka M, Kunze L. Explanations in autonomous driving: A survey. IEEE Trans. on Intelligent Transportation Systems, 2022, 23(8): 10142–10162. [doi: 10.1109/TITS.2021.3122865]

[13] Atakishiyev S, Salameh M, Yao HS, Goebel R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. IEEE Access, 2024, 12: 101603–101625. [doi: 10.1109/ACCESS.2024.3431437]

[14] Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access, 2023, 11: 78994–79015. [doi: 10.1109/ACCESS.2023.3294569]

[15] Li XH, Xiong HY, Li XJ, Wu XY, Zhang X, Liu J, Bian J, Dou DJ. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 2022, 64(12): 3197–3234. [doi: 10.1007/s10115-022-01756-8]

[16] Islam SR, Eberle W, Ghafoor SK, Ahmed M. Explainable artificial intelligence approaches: A survey. arXiv:2101.09429, 2021.

[17] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy, 2020, 23(1): 18. [doi: 10.3390/e23010018]

[18] Yang PB, Sang JT, Zhang B, Feng YG, Yu J. Survey on interpretability of deep models for image classification. Ruan Jian Xue Bao/Journal of Software, 2023, 34(1): 230–254 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6415.htm [doi: 10.13328/j.cnki.jos.006415]

[19] Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 2018, 73: 1–15. [doi: 10.1016/j.dsp.2017.10.011]

[20] Zurada JM, Malinowski A, Cloete I. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In: Proc. of the 1994 IEEE Int'l Symp. on Circuits and Systems. London: IEEE, 1994. 447–450. [doi: 10.1109/ISCAS.1994.409622]

[21] Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: An overview. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019. 193–209. [doi: 10.1007/978-3-030-28954-6_10]

[22] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

[23] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 2668–2677.

[24] Chen CF, Li O, Tao CF, Barnett AJ, Su JK, Rudin C. This looks like that: Deep learning for interpretable image recognition. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 8930–8941.

[25] Yang Q, Fan LX, Zhu J, Chen YX, Zhang QS, Zhu SC, Tao DC, Cui P, Zhou SH, Liu Q, Huang XQ, Zhang YF. Introduction to Explainable Artificial Intelligence. Beijing: Electronics Industry Press, 2022 (in Chinese).

[26] Margolis E. A reassessment of the shift from the classical theory of concepts to prototype theory. Cognition, 1994, 51(1): 73–89. [doi: 10.1016/0010-0277(94)90009-4]

[27] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proc. of the 13th European Conf. on Computer Vision (ECCV 2014). Zurich: Springer, 2014. 818–833. [doi: 10.1007/978-3-319-10590-1_53]

[28] Molnar C. Interpretable Machine Learning. Lulu.com, 2020.

[29] Crabbé J, van der Schaar M. Concept activation regions: A generalized framework for concept-based explanations. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2590–2607.

[30] Zhou BL, Sun YY, Bau D, Torralba A. Interpretable basis decomposition for visual explanation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 122–138. [doi: 10.1007/978-3-030-01237-3_8]

[31] Goyal Y, Feder A, Shalit U, Kim B. Explaining classifiers with causal concept effect (CaCE). arXiv:1907.07165, 2020.

[32] Wu ZX, D'Oosterlinck K, Geiger A, Zur A, Potts C. Causal proxy models for concept-based model explanations. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 37313–37334.

[33] Zhou BL, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene CNNs. arXiv:1412.6856, 2015.

[34] Bau D, Zhou BL, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3319–3327. [doi: 10.1109/CVPR.2017.354]

[35] Fong R, Vedaldi A. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8730–8738. [doi: 10.1109/CVPR.2018.00910]

[36] Ghorbani A, Wexler J, Zou J, Kim B. Towards automatic concept-based explanations. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 9277–9286.

[37] Yeh CK, Kim B, Arik SÖ, Li CL, Pfister T, Ravikumar P. On completeness-aware concept-based explanations in deep neural networks. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 20554–20565.

[38] Zhang RH, Madumal P, Miller T, Ehinger KA, Rubinstein BIP. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 11682–11690. [doi: 10.1609/aaai.v35i13.17389]

[39] Vielhaben J, Blücher S, Strodthoff N. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. arXiv:2301.11911, 2023.

[40] Fel T, Picard A, Bethune L, Boissin T, Vigouroux D, Colin J, Cadène R, Serre T. CRAFT: Concept recursive activation factorization for explainability. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 2711–2721. [doi: 10.1109/CVPR52729.2023.00266]

[41] Yang LJ, Wang JQ, Jing LP, Yu J. Semantic representation learning of convolutional neural network based on tensor computation. Chinese Journal of Computers, 2023, 46(3): 568–578 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2023.00568]

[42] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 2019, 1(5): 206–215. [doi: 10.1038/s42256-019-0048-x]

[43] Zhai JH, Zhang SF, Chen JF, He Q. Autoencoder and its various variants. In: Proc. of the 2018 IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC). Miyazaki: IEEE, 2018. 415–419. [doi: 10.1109/SMC.2018.00080]

[44] Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114, 2022.

[45] Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, Liang P. Concept bottleneck models. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 5338–5348.

[46] Kim E, Jung D, Park S, Kim S, Yoon S. Probabilistic concept bottleneck models. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 16521–16540.

[47] Chen Z, Bei YJ, Rudin C. Concept whitening for interpretable image recognition. Nature Machine Intelligence, 2020, 2(12): 772–782. [doi: 10.1038/s42256-020-00265-z]

[48] Zarlenga ME, Barbiero P, Ciravegna G, Marra G, Giannini F, Diligenti M, Shams Z, Precioso F, Melacci S, Weller A, Lio P, Jamnik M. Concept embedding models: Beyond the accuracy-explainability trade-off. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 21400–21413.

[49] Yuksekgonul M, Wang M, Zou J. Post-hoc concept bottleneck models. arXiv:2205.15480, 2023.

[50] Rigotti M, Miksovic C, Giurgiu I, Gschwind T, Scotton P. Attention-based interpretability with concept Transformers. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[51] Zhang QS, Wu YN, Zhu SC. Interpretable convolutional neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8827–8836. [doi: 10.1109/CVPR.2018.00920]

[52] Alvarez-Melis D, Jaakkola TS. Towards robust interpretability with self-explaining neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 7786–7795.

[53] Wang BW, Li LZ, Nakashima Y, Nagahara H. Learning bottleneck concepts in image classification. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 10962–10971. [doi: 10.1109/CVPR52729.2023.01055]

[54] Rajagopal D, Balachandran V, Hovy EH, Tsvetkov Y. SelfExplain: A self-explaining architecture for neural text classifiers. In: Proc. of

the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 836–850. [doi: 10.18653/v1/2021.emnlp-main.64]

[55] Nauta M, van Bree R, Seifert C. Neural prototype trees for interpretable fine-grained image recognition. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14928–14938. [doi: 10.1109/CVPR46437.2021.01469]

[56] Hase P, Chen CF, Li O, Rudin C. Interpretable image recognition with hierarchical prototypes. In: Proc. of the 7th AAAI Conf. on Human Computation and Crowdsourcing. Washington: AAAI, 2019. 32–40. [doi: 10.1609/hcomp.v7i1.5265]

[57] Wang JQ, Liu HF, Wang XY, Jing LP. Interpretable image recognition by constructing transparent embedding space. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 875–884. [doi: 10.1109/ICCV48922.2021.00093]

[58] Wang JQ, Liu HF, Jing LP. Transparent embedding space for interpretable image recognition. IEEE Trans. on Circuits and Systems for Video Technology, 2024, 34(5): 3204–3219. [doi: 10.1109/TCSVT.2023.3314769]

[59] Rymarczyk D, Struski Ł, Tabor J, Zieliński B. ProtoPShare: Prototypical parts sharing for similarity discovery in interpretable image classification. In: Proc. of the 27th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining. ACM, 2021. 1420–1430. [doi: 10.1145/3447548.3467245]

[60] Xue MQ, Huang QH, Zhang HF, Cheng LC, Song J, Wu MH, Song ML. ProtoPFormer: Concentrating on prototypical parts in vision Transformers for interpretable image recognition. arXiv:2208.10431, 2022.

[61] Donnelly J, Barnett AJ, Chen CF. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10255–10265. [doi: 10.1109/CVPR52688.2022.01002]

[62] Wang BS, Wang CY, Chiu WC. MCPNet: An interpretable classifier via multi-level concept prototypes. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 10885–10894. [doi: 10.1109/CVPR52733.2024.01035]

[63] Sawada Y, Nakamura K. Concept bottleneck model with additional unsupervised concepts. IEEE Access, 2022, 10: 41758–41765. [doi: 10.1109/ACCESS.2022.3167702]

[64] Sarkar A, Vijaykeerthy D, Sarkar A, Balasubramanian VN. A framework for learning ante-hoc explainable models via concepts. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10276–10285. [doi: 10.1109/CVPR52688.2022.01004]

[65] Marconato E, Passerini A, Teso S. GlanceNets: Interpretable, leak-proof concept-based models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 21212–21227.

[66] Yang Y, Panagopoulou A, Zhou SH, Jin D, Callison-Burch C, Yatskar M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 19187–19197. [doi: 10.1109/CVPR52729.2023.01839]

[67] Oikarinen T, Das S, Nguyen LM, Weng TW. Label-free concept bottleneck models. arXiv:2304.06129, 2023.

[68] Wang JQ, Wang PC, Feng Y, Liu HF, Gao C, Jing LP. Align2Concept: Language guided interpretable image recognition by visual prototype and textual concept alignment. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 8972–8981. [doi: 10.1145/3664647.3680707]

[69] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[70] Colin J, Fel T, Cadène R, Serre T. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2832–2845.

[71] Rosch E. Cognitive representations of semantic categories. Journal of Experimental Psychology: General, 1975, 104(3): 192–233. [doi: 10.1037/0096-3445.104.3.192]

[72] Gärdenfors P. Conceptual Spaces: The Geometry of Thought. Cambridge: MIT Press, 2000.

[73] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 24824–24837.

[74] Marvin G, Hellen N, Jjingo D, Nakatumba-Nabende J. Prompt engineering in large language models. In: Proc. of the 2024 Int'l Conf. on Data Intelligence and Cognitive Informatics. Singapore: Springer, 2024. 387–402. [doi: 10.1007/978-981-99-7962-2_30]

[75] Zhao HY, Chen HJ, Yang F, Liu NH, Deng HQ, Cai HY, Wang SQ, Yin DW, Du MN. Explainability for large language models: A survey. ACM Trans. on Intelligent Systems and Technology, 2024, 15(2): 20. [doi: 10.1145/3639372]

[76] LCM Team, Barrault L, Duquenne PA, Elbayad M, Kozhevnikov A, Alastruey B, Andrews P, Coria M, Couairon G, Costa-Jussà MR, Dale D, Elsahar H, Heffernan K, Janeiro JM, Tran T, Ropers C, Sánchez E, Roman RS, Mourachko A, Saleem S, Schwenk H. Large concept models: Language modeling in a sentence representation space. arXiv:2412.08821, 2024.

## 附中文参考文献

[18] 杨朋波, 桑基韬, 张彪, 冯耀功, 于剑. 面向图像分类的深度模型可解释性研究综述. 软件学报, 2023, 34(1): 230–254. http://www.jos.org.cn/1000-9825/6415.htm [doi: 10.13328/j.cnki.jos.006415]

[25] 杨强, 范力欣, 朱军, 陈一昕, 张拳石, 朱松纯, 陶大程, 崔鹏, 周少华, 刘琦, 黄萱菁, 张永锋. 可解释人工智能导论. 北京: 电子工业出版社, 2022.

[41] 杨礼吉, 王家祺, 景丽萍, 于剑. 基于张量计算的卷积神经网络语义表示学习. 计算机学报, 2023, 46(3): 568–578. [doi: 10.11897/SP.J.1016.2023.00568]

## 作者简介

**王家祺**, 博士, 主要研究领域为深度学习可解释性, 计算机视觉.

**冯毅**, 博士生, CCF 学生会员, 主要研究领域为自然语言处理, 情感计算, 信息抽取.

**刘华锋**, 博士, 讲师, CCF 专业会员, 主要研究领域为机器学习, 智能推荐.

**景丽萍**, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 高维数据表示及应用.

**于剑**, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为人工智能, 机器学习.