面向零样本图像分类的交互式类属性构建方法

刘真, 徐景胜, 颜菁, 徐润森, 吴向阳*

(杭州电子科技大学计算机学院 杭州 310018)
(wuxy@hdu.edu.cn)

摘 要: 零样本图像分类解决了训练和测试数据类别不相交的问题, 人类标注属性是一种常用的实现零样本图像分类的辅助知识. 为协助专家设计类属性矩阵, 提出了一种交互式构建方法, 简化了烦琐且缺乏指导的流程. 首先, 通过一种基于概念的深度学习可解释性方法, 在训练集图像数据中提取出可理解的属性信息; 然后, 采用多视图协作的交互方式, 探索和分析已提取属性的重要性. 系统提供了全局和局部 2 种方式, 辅助用户设计测试集数据类别的属性值; 最后, 通过在数据集 Animals with Attributes2 上进行的案例分析, 以及采用李克特量表的用户评估实验, 验证了设计方法的有效性和实用性, 可以帮助专家用户高效且便捷地完成类属性构建工作.

关键词: 零样本学习; 零样本图像分类; 可视分析; 可解释人工智能; 人机协作
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00349

Interactive Class Attribute Construction Method for Zero-Shot Image Classification

Liu Zhen, Xu Jingsheng, Yan Jing, Xu Runsen, and Wu Xiangyang*

(*School of Computer Science, Hangzhou Dianzi University, Hangzhou* 310018)

Abstract: Zero-shot image classification addresses the problem of non-overlapping categories between training and testing data. Human-annotated attributes are a commonly used form of auxiliary knowledge in zero-shot image classification. To assist experts in designing class attribute matrices, an interactive construction method is proposed, simplifying the tedious and unguided process. First, understandable attribute information is extracted from the training dataset images using a concept-based deep learning explainability method. Then, an interactive approach involving multi-view collaboration is employed to explore and analyze the importance of the extracted attributes. The system offers both global and local options to aid users in designing attribute values for testing dataset categories. Finally, the effectiveness and practicality of the design method are validated through case studies on the Animals with Attributes2 dataset and user evaluation experiments using the Likert scale, facilitating efficient and convenient class attribute construction for expert users.

Key words: zero-shot learning; zero-shot image classification; visual analysis; explainable artificial intelligence; human-machine collaboration

一般的有监督机器学习需要大量的标记数据用于模型训练, 并且只能对包含在训练数据类别中的测试实例进行有效的预测分类. 零样本图像分类问题中, 出现在训练样本中的类别称为已见

收稿日期: 2023-06-18; 修回日期: 2023-11-27. **基金项目**: 浙江省自然科学基金(LY22F020023); 浙江省自然科学基金重点项目(LZ22F020015); 国家自然科学基金面上项目(61972122). **刘真**(1977—), 女, 博士, 副教授, 硕士生导师, 主要研究方向为数据可视化与可视分析、深度学习; **徐景胜**(1999—), 男, 硕士研究生, 主要研究方向为数据可视化与可视分析; **颜菁**(1998—), 女, 硕士研究生, 主要研究方向为数据可视化与可视分析; **徐润森**(2000—), 男, 硕士研究生, 主要研究方向为数据可视化与可视分析; **吴向阳**(1975—), 男, 博士, 教授, 硕士生导师, CCF 会员, 论文通信作者, 主要研究方向为数据可视化与可视分析、计算机图形学.

类, 在训练期间未出现类别称为未见类. 在人类推理过程中, 一个从未见过斑马的人可以通过长得有黑白条纹的马来识别斑马. 零样本图像分类的思想与此类似, 仅使用已见类的数据和物体的描述进行模型训练, 就能够正确地识别未见类的图像实例[1-2]. 由于其更符合实际应用场景, 并减少了对数据集的依赖, 因此在最新的机器学习和计算机视觉领域中取得广泛应用.

实现零样本图像分类的一个关键要素是引入外部知识(也称辅助信息), 如文本描述、图像注释和词向量等[3-5], 通过这些信息来建立已见类与未见类之间的联系, 使训练完成的模型可以对输入的未见类实例数据进行有效的过滤和提取, 并获得最终正确的分类结果. 其中, 人工定义的属性是一种广泛使用的辅助信息, 用于零样本图像分类的属性是由领域专家设计的人类可理解的语义词, 如黑色和花斑等. 专家通过分析已见类和未见类是否具有或有多大可能具有这个属性特征, 为每个类的各属性设置一个值. 整个设计过程中存在许多困难和问题, 首先, 需要花费大量的时间和精力来设计所有类别共享的属性集合. 由于各个数据集包含的图像类别不同, 在某个数据集中适用的属性集合在另一个数据集中不能完全适用, 因此对每个数据集都需要重新设计一个符合当前数据集要求的属性集合[1,6-7]. 其次, 设计好一个属性集合之后, 领域专家还需要标记出类别中是否存在这些属性代表的特征, 或是出现这些属性特征的可能性. 在数据集的类别数量和属性集合的数量都非常庞大的情况下, 这将是一项艰巨的工作[8]. 最后, 只有在设计工作全部完成之后, 才能评估设计的类属性矩阵是否满足任务需求, 但也无法判断属性对分类结果具有积极还是消极的作用.

为了降低类属性设计的难度, 有工作提出了一种视觉可解释的主动学习方法, 通过 4 个动作(提问、解释、推荐、反馈)优化人类与人工智能的协作, 提高了人类交互式构建零样本分类模型的效率[9], 但这仍然需要系统使用者参与设计的全部过程; 还有通过设计独特的神经网络模块提取类关系, 以实现几乎完全不需要人工参与的类属性矩阵设计[10], 然而该方法在很大程度上依赖于词向量的选择.

为了解决上述问题, 本文面向零样本图像分类问题, 提出了一种交互式类属性矩阵构建的可视分析方法. 它旨在综合运用人工智能和可视化技术, 帮助专家用户高效和便捷地完成类属性构建. 主要贡献如下:

(1) 提出一种类属性生成方法, 该方法基于概念的深度学习可解释性思想, 支持从图像中提取类别属性, 通过一组图像特征块提供不同属性的语义可解释性, 并支持对不同类别的各维属性值的局部调整.

(2) 设计并实现了一个实时交互属性设计系统, 它支持快速识别和选择重要性较高的属性, 提供了全局和局部 2 种高效交互方式.

(3) 通过案例研究和用户反馈定性地评估了本系统的有效性、学习难度和便捷性. 与没有指导的设计过程相比, 本系统可以显著地提高专家用户属性构建效率并改善设计体验.

1 相关工作

1.1 零样本学习

零样本学习旨在对未见类图像实例进行分类. 现有的零样本学习方法通常假设已见类和未见类共享一个公共语义空间, 因此实现零样本学习的关键思想是将知识从已见类转移到未见类. 目前零样本学习方法可以分为基于嵌入模型和基于视觉样本生成 2 大类[11]. 基于嵌入模型的方法期望学习一个从视觉空间到语义空间的映射, 将未见类的实例投射到属性嵌入空间, 然后通过寻找最相似的类属性来预测它们的类标签[12-15]. Kodirov 等[16]提出应用编码器-解码器结构, 在重建损失的监督下提取更多特征信息. Changpinyo 等[17]提出将语义空间和图像空间对齐, 从而在嵌入中提取更多的语义信息. 生成式方法在根据属性信息合成实例的基础上对未见类进行分类[18-20], Zhu 等[20]和 Xian 等[21]均是利用语义信息来生成卷积神经网络(convolutional neural network, CNN)特征, 并利用瓦瑟斯坦生成对抗网络(Wasserstein generative adversarial network, WGAN)和分类损失, 生成判别性强的 CNN 特征训练分类器. Verma 等[22]采用条件自动编码器[23]生成可见类和不可见类的特征, 以训练分类器.

1.2 视觉属性设计

作为零样本学习的一个重要部分, 属性可以将具有共同特征的不同类别联系起来, 但与零样本学习方法的研究相比, 视觉属性设计的探索相对不足. 通常, 视觉属性的设计是通过人工挑选一组对图像具有描述性的词[1,6]实现的, 这个过程需

要特殊的专业知识和有关领域的专业人士. 为了减小工作量，之前也有工作探索了一些自动化的方法辅助设计，如 Bergamo 等[24]通过学习二进制代码来探索图像级嵌入，Berg 等[25]提出了一种通过挖掘从互联网采样的文本和图像数据来处理视觉属性标记图像的自动方法. 然而，从文本挖掘中发现的属性可能包含与当前的分类任务无关的语义，或者在视觉特征空间中或许不可分离. Parikh 等[26]通过在算法迭代中引入人类操作来发现局部的和可命名的属性. 基于此，Duan 等[27]提出了一种发现细粒度识别的局部属性的方法. Yu 等[7]提出了一种自动方法来设计有区别的类别级属性. 但是，学习到的属性不会像手动指定的属性那样具有简洁的语义. 此外，对于未见类，这种方法需要人类对已见类和未见类之间的相似性进行建模，这比直接指定未见类的视觉属性更加困难和不直观. Visual transformer[28]和 BagNets[29]表明，图像块可以作为强大的视觉词来传达用于类别区分的视觉线索.

1.3　基于概念的深度学习可解释性

使用概念来提高深度学习的可解释性有 2 种主要方法[30]，用基于概念的约束来训练已有的可解释性模型和构建基于概念的事后解释. 关于第 1 种方法，Koh 等[31]提出了概念瓶颈模型(concept bottleneck models, CBM)，该模型限制神经网络，使其以自动编码器的方式运行；其将输入映射到人类可解释的概念，并基于这些概念进行预测. Caruana 等[32]提出了 2 个案例研究，在这 2 个案例中，具有成对相互作用的高性能广义加性模型(generative adversarial active models, GA2Ms)被应用于真实的医疗问题，生成了具有较高精度的可理解模型. 虽然这些方法采用概念来提供内嵌的可解释性，但它们均未提供深入了解已经训练完成的模型的方案. 对于第 2 种方法，通常是一种基于扰动/敏感性分析的可解释性方法. Koh 等[33]使用数据点和特征作为扰动的形式，并分析模型的输出如何变化. Kim 等[34]提供了一种称为使用概念激活向量进行测试(testing with concept activation vectors, TCAV)的方法来定义一个概念，并计算其对模型预测的特定类别影响(称为 TCAV 分数). 在 TCAV 的基础上，Ghorbani 等[35]提出基于概念的自动解释(automatic concept-based explanations, ACE)，是一种从数据集中自动提取有影响力的特定类别概念的技术. Ge 等[36]提出了视觉推理解释(visual reasoning explanation, VRX)框架，其通过概念回答可解释性问题.

2　系统概述

2.1　设计目标

针对相关零样本学习属性研究面临的挑战进行调查分析，总结出本系统的主要目标如下:

G1. 展示算法提取的各属性值在不同类别上的分布情况. 类别的各维度属性值是本文研究的重点，在视图中展示各维属性在不同类别中的分布趋势，可以加深用户对属性具体含义的理解.

G2. 展示提取的各维度属性表示的含义. 需要局部设计未见类别各个维度的属性值时，必须了解该属性的含义，因此需要通过某种方式展示每个属性代表的含义.

G3. 探究算法提取的各维属性的重要程度. 通过算法生成的属性的重要性会直接影响零样本学习的结果，较低重要性的属性可能会对模型的训练产生负面影响，因此展示生成的属性的重要性程度也是优化属性必不可少的一部分.

G4. 提供简捷高效的交互方式探索未见类的属性值. 未见类属性需要用户单独设计，这是本系统中人机协作的重要部分，因此，需要提供一套便捷高效的操作方法，实现用户对属性值的交互式探索和分析.

2.2　系统框架

基于以上设计目标，本文设计的可视化系统由后端数据处理和前端数据可视化 2 部分组成，系统流程如图 1 所示. 后端数据处理部分在已见类数据中提取属性信息，随后在前端界面的各视图中展示不同信息，通过可视交互与用户协作.

2.2.1　问题描述

本节简要介绍零样本分类的基本概念，并阐述研究问题.

在零样本图像分类中的训练数据和测试数据包含的类别分别称为已见类和未见类，分别定义为 $C_{\mathrm{s}}=\left\{c_i^{\mathrm{s}} \mid i=1,2,\cdots,N_{\mathrm{s}}\right\}$ 和 $C_{\mathrm{u}}=\left\{c_i^{\mathrm{u}} \mid i=1,2,\cdots,N_{\mathrm{u}}\right\}$；它们是不相交的，即 $C_{\mathrm{s}}\bigcap C_{\mathrm{u}}=\varnothing$. 训练数据和测试数据都来自表示一个称为 $\boldsymbol{\mathcal{X}}\in\mathbb{R}^D$ 的 D 维特征空间，定义训练数据集为 $D_{\mathrm{tr}}=\left\{\left(x_i^{\mathrm{tr}},y_i^{\mathrm{tr}}\right)\in\boldsymbol{\mathcal{X}}\times C_{\mathrm{s}}\right\}$. 其中，$x_i^{\mathrm{tr}}$ 表示特征空间中的有标签实例；y_i^{tr} 表示已见类中对应的类标签. 类似地，将 $D_{\mathrm{te}}=\left\{x_i^{\mathrm{te}}\in\boldsymbol{\mathcal{X}}\right\}$ 表示为测试数据集，其中 x_i^{te} 表示特征空间中无标签的实例. 给定数据集 D_{tr}，零样本分类旨在学习一个分类模型 $f(\cdot):\boldsymbol{\mathcal{X}}\to C_{\mathrm{u}}$ 来预测测试数据集 D_{te} 的标签[9].

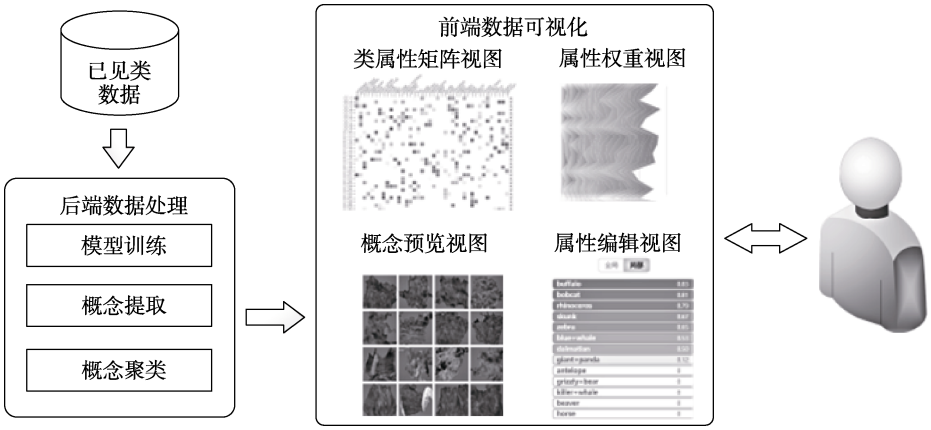


图 1 系统流程图

由于在训练期间没有未见类的实例，模型需要一些辅助信息来将知识从已见类转移到未见类. 最广泛使用的辅助信息是属性，所有属性构成了一个空间称为语义空间 $\boldsymbol{\mathcal{A}} \in \mathbb{R}^M\ (M < D)$，空间中每个维度代表一个属性，属性值可以是二进制 0 和 1，表示这个属性在当前类别中是否存在，也可以是连续值来表示置信度. 在属性语义空间中，每个类都表示为一个语义向量，被称为类原型；将已见类的类原型表示为 $\boldsymbol{\mathcal{A}}_{\mathrm{s}} = \left\{a_i^{\mathrm{s}} \mid a_i^{\mathrm{s}} \in \boldsymbol{\mathcal{A}}\right\}_{i=1}^{N_{\mathrm{s}}}$，未见类的类原型 $\boldsymbol{\mathcal{A}}_{\mathrm{u}} = \left\{a_i^{\mathrm{u}} \mid a_i^{\mathrm{u}} \in \boldsymbol{\mathcal{A}}\right\}_{i=1}^{N_{\mathrm{u}}}$. 将所有的类原型垂直对齐即可得到类属性矩阵，本文旨在交互式地构建类属性矩阵.

2.2.2 TCAV

TCAV 是一种基于概念的解释机器学习模型方法. 该方法旨在帮助人们理解深度学习模型的决策过程，通过对模型内部的激活向量进行解释，揭示模型对于不同特征和概念的响应.

使用 TCAV，首先需要定义一个待分析的概念. 可以通过一组包含该概念的图像实例，或者一个带有概念标签的独立数据集来表示这个概念，另外还需要选择一组不包含这个概念的图像示例作为反例. 其次，通过一个训练完成的深度学习模型获取概念图像的高维表示，即获取所有数据在模型某一层的激活值. 然后以概念示例的高维表示与非概念示例的高维表示为 2 种类别来训练一个二元线性分类器，TCAV 将概念激活向量(concept activation vectors, CAV)定义为分类器超平面的法线，该超平面将概念示例的高维表示与非概念示例的高维表示线性分开.

为了衡量概念对实例预测的影响，TCAV 计算实例中间层激活值在 CAV 方向上对模型输出层影响的方向导数. “概念敏感性”通过这个导数来表示，导数值大于零即意味着概念对预测有着积极的影响；反之，意味着概念对预测有负面的影响. 最后，使用上述方式对类别中的所有实例进行测试，计算出每个实例的概念敏感性值. 将概念敏感性值大于零的实例数量与类中所有实例的数量的比值，作为模型“认为”这个用户选择的概念对当前类别分类的重要性程度.

2.2.3 ACE

TCAV 提供了一种机制来衡量人类可理解的概念对特定类别的深度学习模型的影响，但在该方法中为一个概念设置一组例子是一个人工过程. 为了使概念的生成和提取过程自动化，Ghorbani 等[35]提出了 ACE.

ACE 方法首先对各类别的图像进行分割，并在模型中间层获取图像块的高维空间表示，再将这些高维空间表示的聚类结果作为每个类别一组概念，其中可以根据一些规则去除部分聚类簇. 最后遵循 TCAV 方法获取这些自动提取概念的 TCAV 分数. 在本系统中，默认每个类别随机采样至少 50 幅、至多 200 幅图像，并使用 SLIC 对图像进行分割. 参照 ACE 文章的工作，分别将每幅图像分割为 15，50 和 80 块，使用 *K*-means 算法聚类它们的高维表示.

2.2.4 类属性生成

采用 TCAV 和 ACE 方法，可以为每一个已见类提取一组概念，但每个已见类的概念组是单独的特征空间，不能直接作为属性来使用. 因为一个属性应该是由所有类别共享的变量，而目前每个概念只属于自己的类别. Huang 等[30]指出，概念可以在不同的和看似不同的类之间共享. 例如，在斑马类别中提取到的斑纹概念和在狮子鱼类别中提

取到的斑纹概念在视觉上具有较高的相似度. 这表明某些概念在不同类别中都存在，本文认为这些跨类别存在的概念是广泛存在的属性. 因此，将所有已见类的概念也进行了聚类操作，借此将语义相似的概念分组到同一个集群中，把一个概念集群作为一个属性. 一个类别的属性值用概念的 TCAV 值表示，当一个类别的所有概念都不在某个概念集群中出现，则将此类别的属性值设置为 0.

在后端处理部分，本文先选择一个主流的神经网络模型 GoogleNet，并在训练数据集上进行训练. 遵循 ACE 方法的流程，对于每个已见类，提取到一组概念集合. 随后将所有类别的概念进行聚类操作，每一个概念集群作为一个属性，概念的 TCAV 值作为属性值，这样可以获得已见类的类属性矩阵.

3　可视化设计

3.1　基础参数设置模块

3.1.1　配置面板

配置面板区域用于设置整个系统的可选参数，如图 2 中 A 区域所示. 它是整个可视化系统的入口，可以在此选择深度学习模型、数据集和提取高维特征的中间层，设置完成后整个系统随参数展示出各个视图.

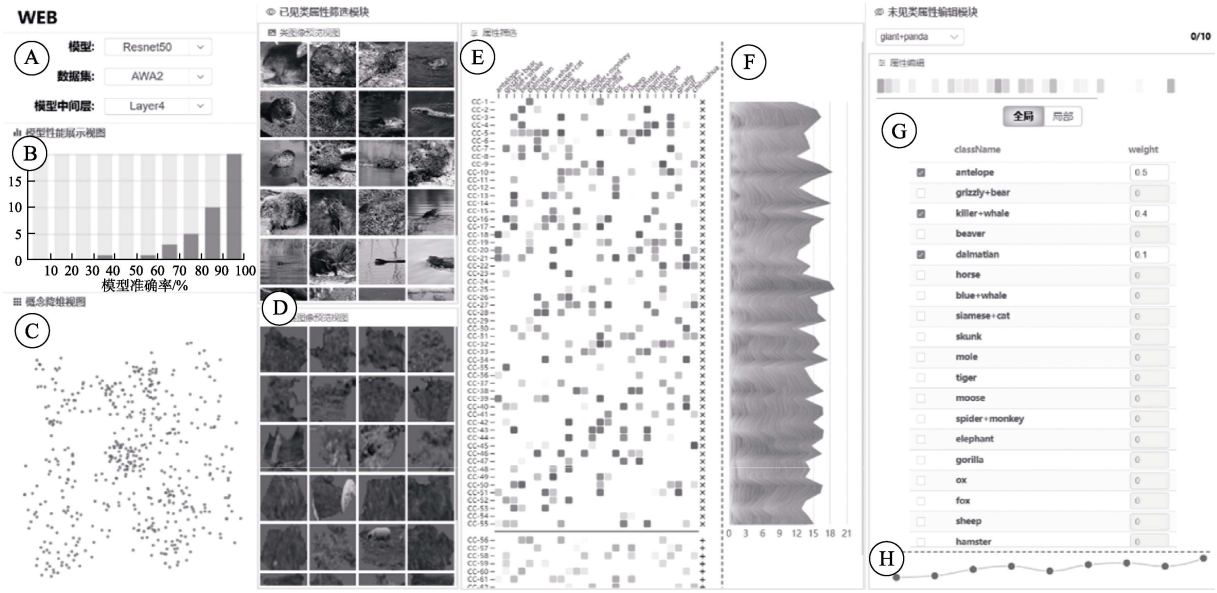


图 2　系统界面图

3.1.2　模型性能视图

模型性能视图展示参数选择的数据集在深度学习模型上的分类精度信息，如图 2 中区域 B 所示. 各个类别的分类正确率的分布通过柱状图的高度显示，例如，x 轴 70~80 的柱形高度代表着分类正确率为 70%~80%的类别数量. 具有较高分类正确率的模型可以更好地提取各个类别的概念，且对最后生成的类属性矩阵有直接影响. 此视图仅展示已见类的信息，由于遵循零样本分类的设定，因此在对零样本分类模型进行测试之前，无法获得未见类的图像数据.

3.1.3　概念降维视图

概念降维视图采用散点图形式将所有概念展示在图 2 区域 C 中. 概念是由一组图像块集合表示的，因此无法直接进行降维和聚类操作. 本文首先将概念图像块输入神经网络模型，将中间层的激活输出作为概念的高维特征. 一个概念的高维特征用属于它的所有图像块的高维特征的平均值来表示，通过这个操作将每个概念转化为一个高维向量，然后使用 t-SNE 算法[37]对概念数据进行降维处理.

3.2　已见类属性筛选模块

3.2.1　已见类属性矩阵视图

为了直观地展示出已见类的属性分布，已见类属性矩阵视图以类属性矩阵的形式将所有属性值展示在图 2 区域 E 中. 属性值与 TCAV 分数相对应，数值范围为 0~1. 属性值越大，在视图上显示的单元格颜色越深；相反，在视图上显示的单元格颜色则越浅. 每一行表示一个属性的属性值，左侧显示的是属性的编号，右侧显示“删除属性”操作的选项，可以通过属性权重堆叠图去除重要

性程度较低的一些属性，在类属性矩阵的下方展示了已经去除的属性行；某些情况下可以通过点击右侧的“添加”操作选项重新选择属性. 类属性矩阵的每一列展示一个类的所有属性，上方展示类别的名称.

3.2.2 属性权重堆叠视图

属性权重图使用堆叠折线图将属性权重值展示在图 2 区域 F 中. 为了满足系统实时交互的需求，本文采用线性回归方法为基础，分析各个属性的重要性程度. 相比于深度学习技术，传统机器学习方法所需的训练时间更少，能够根据用户的操作及时地给出反馈结果.

具体地，首先，定义类视觉原型为一个类中所有实例的高维特征的平均. 与获取概念的高维特征的方式相同，将实例输入神经网络模型中，模型中间层的激活输出作为该实例的高维特征，每一个类可以转化为一个 d 维的类原型向量. 给定类视觉原型矩阵和类属性矩阵，学习 d 个线性回归模型，并对应类原型的每一个维度. 线性回归模型 $f_i:\boldsymbol{W}_i\boldsymbol{X}+\boldsymbol{b}_i, i=1,2,\cdots,d$ 的权重 $\boldsymbol{W}$ 与类属性向量具有相同的维度，权重 $\boldsymbol{W}$ 的每一维值代表着输入的类属性值对输出的预测值的重要程度. 假设在所有的线性回归模型中某一维类属性对应的权重值都较低，则表示这一维类属性对最后的预测结果都有较小的影响，可以认为该维属性具有较低的重要性，分析其原因可能是概念提取效果不佳或是概念聚类效果不佳.

本文使用堆叠折线图把线性回归模型的权重值展示到视图中，用户可以非常直观地看出哪一维类属性矩阵被赋予了最低的权重值. 结合已见类属性矩阵图选择是否需要删除属性，当有属性被用户选择删除之后，系统会重新训练每一个线性回归模型，属性权重堆叠图也随之重新绘制，通过迭代操作可以选择出最优的已见类属性矩阵.

3.2.3 属性图像和类图像预览视图

属性图像和类图像预览视图用于展示属性图像块和类图像实例，界面如图 2 区域 D 所示. 与一般的类属性矩阵设计不同，本文中的属性是由概念集群表示的，无法用概括性的词来说明. 因此，本文系统提供了一个图像展示视图，用于显示各个属性的图像块. 通过点击已见类属性矩阵图中的属性编号来选择展示不同的类属性，用户通过直接观察代表这个属性的图像块，可以得到这个概念的抽象认识. 例如，一个属性展示的图像块大都是颜色相间的条纹，代表着这个属性应该是类似花斑的含义. 另外，也提供了一个用于展示类别的实例图像的视图，当用户需要了解类的具体信息时，可以通过点击已见类属性矩阵图中的类别名称选择不同的类进行展示.

3.3 未见类属性编辑模块

通过已见类属性筛选模块的各个功能，用户可以在算法生成的所有属性中筛选出包含重要信息的一组属性. 因为根据零样本学习的设定，只能获得未见类的类别名而没有未见类的图像数据可以使用，需要借助用户的知识编辑出未见类属性的类属性矩阵.

使用本模块第 1 步需要在下拉列表中选择一个待编辑的未见类，每一个未见类的类属性值都默认设置为 0，系统提供了全局和局部 2 种方式对未见类的类属性向量进行编辑.

3.3.1 全局编辑视图

全局编辑视图提供交互视图帮助用户在全局层面上设计未见类属性值，界面展示在图 2 区域 G 中. 设计全局编辑方式编辑未见类属性向量是通过组合已见类的类属性向量来实现的. 首先，需要用户选择出若干个在视觉上与当前待编辑的未见类较相似的已见类，并单独设置每个已见类的权重值，随后按照权重值将已见类的类属性向量进行组合，得到当前未见类的类属性向量. 例如，当需要编辑熊猫的类属性向量时，可以选择斑马、北极熊和棕熊为较相似的已见类，再通过设置它们的权重组合获得熊猫的类属性向量.

3.3.2 局部编辑视图

局部编辑视图提供交互视图帮助用户在局部层面上设计未见类属性值. 局部编辑方式编辑未见类属性向量通过单独设置每一维向量的方式来实现. 本文中使用的属性的含义可以通过查看属性图像来了解，但与直接使用描述词来表示属性的方式相比，从图像中获取含义有时并不直观，并且本文使用的属性为 0~1 的连续值，与 0 和 1 的二元属性值相比，在设置上有更多的灵活性，这也会加大设置的难度. 因此，本系统采用排序的方式来设置属性值，如图 3 所示. 在用户选择某一维属性后，将已见类属性值和当前待编辑未见类的这一维属性按属性值降序排列. 用户不需要设置具体的属性值，而是通过拖拽标签块将未见类的属性拖至适当的位置，系统自动设置其值为前后类别属性值的中间值.



图 3　局部数据编辑视图

3.3.3　属性间距折线视图

属性间距折线视图通过折线图展示属性之间的间距，如图 2 区域 H 所示. 在用户每一次编辑操作之后，通过展示当前编辑的未见类的属性与其余各已见类属性的欧几里得距离的平均值，对本次的操作的结果给出反馈. 用户设计出的未见类属性与其他属性相似程度高，使零样本学习模型无法学习到各属性包含的真实信息，这导致最后的分类结果效果不佳. 不同的类别具有相似的类属性向量，也代表着设计的类属性矩阵不具备区分不同类别的能力，不能作为零样本学习有效的辅助知识来使用.

4　实验与评估

4.1　案例分析

使用 AWA2(animals with attributes2)数据集[5]的案例进行实验，以验证本文方法的有效性. AWA2 数据集中包含 50 种哺乳动物的 37 322 幅图像，使用数据集的标准分割(40 个训练类和 10 个测试类)进行实验，选择预训练的 ResNet50 模型[38]作为系统使用的神经网络模型. 此外，AWA2 也提供了一个具有 85 个二元属性的类属性矩阵，该矩阵最初是在一个基于属性的对象相似度心理测试[39]中收集得到的.

实验第 1 步在配置面板中选择本次实验的参数信息. 如图 4 所示，分别选择模型为 ResNet50，数据集为 AWA2，模型中间层为 Layer4. 通过模型性能图可以看出，当前选择的模型在数据集 AWA2 的训练集上整体平均分类正确率较高，因此通过算法提取的属性能够包含丰富的信息，如图 4 所示. 通过后端数据处理后，在 40 个已见类中获得了共 809 个概念，通过聚类算法之后获得共 127 个概念集群，以一定规则去除较小的概念集群后，剩余 104 个概念集群作为后续的属性使用.

第 2 步通过已见类属性筛选模块中的功能筛选过滤掉不含有效信息的属性. 首先，通过浏览图 5 已见类属性矩阵图可以看出其中存在较多空白单元格，即有较多属性值为 0. 数据集提供的类属性矩阵也存在这种现象，这代表着每一个属性都提取到了不同类别图像之间细微的差异，因为一

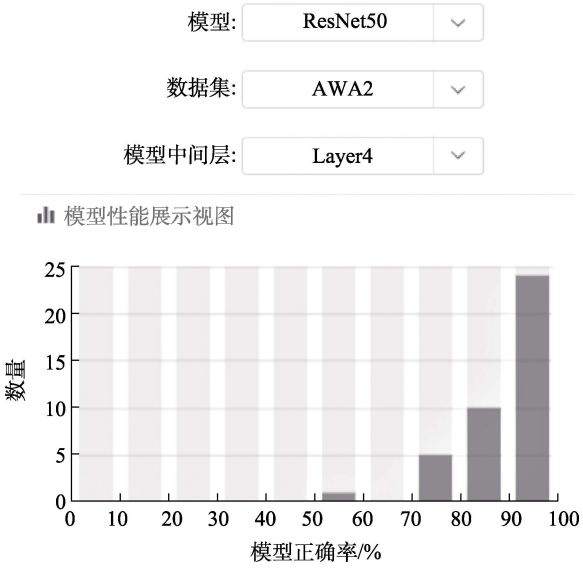


图 4　模型正确率展示

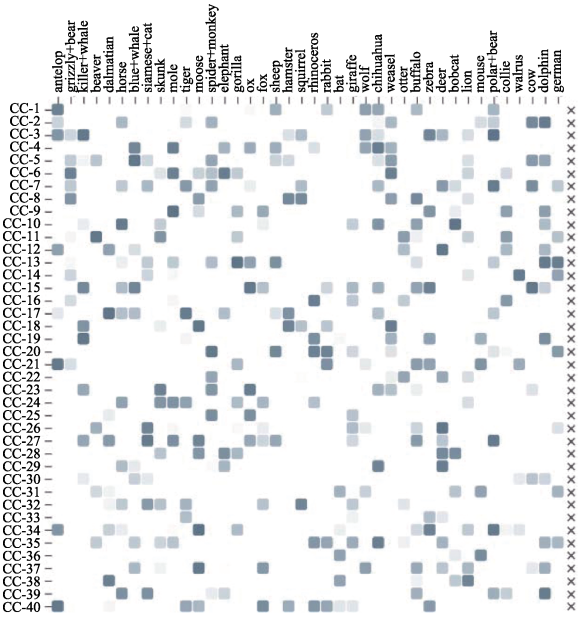


图 5　已见类属性矩阵图

些细小的特征一般只在个别类别中存在，而不会广泛出现在所有类别中. 例如，猫科动物的尖耳朵特征只在 AWA2 数据集训练类中的 5 个类别中出现. 随后，再通过查看图 6 属性权重堆叠图发现，第 26 个属性的权重之和最小. 在图 7 所示属性图像展示区域查看属于这个属性的图像，可知这个属性权重偏低的原因可能是算法并没有很好地提取出这个概念集群的图像，也可能是使用的分割算法效果不佳. 因此在已见类属性矩阵图中去除该属性，属性权重堆叠图随后端模型训练而变化，重复此步骤若干次后去除 6 个属性.

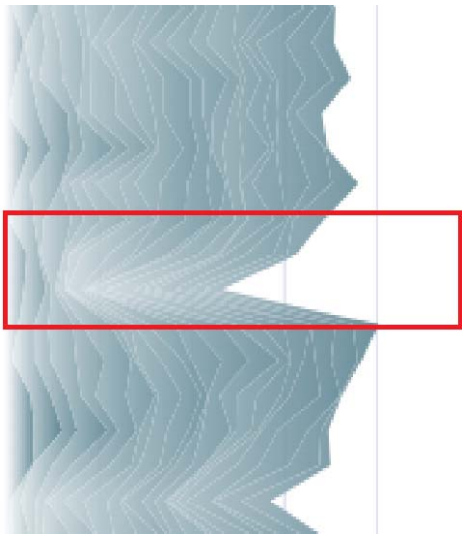


图 6 属性权重值展示

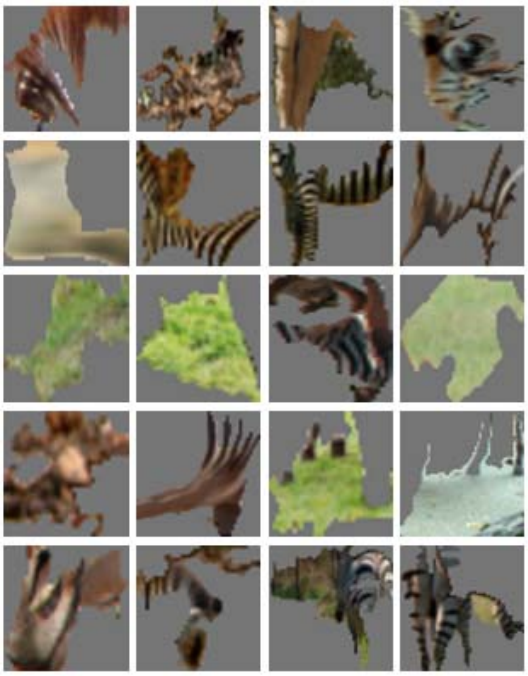


图 7 编号 26 属性块展示

第 3 步使用未见类属性编辑模块的功能视图编辑 10 个未见类的属性向量. 首先，选择待编辑的类别 giant+panda，在全局编辑图中选择与该类别最为相似的若干个已见类，选择了 zebra，polar+bear 和 grizzly+bear 类别，分别设置权重 0.3，0.3 和 0.4，设置方式如图 8 所示.

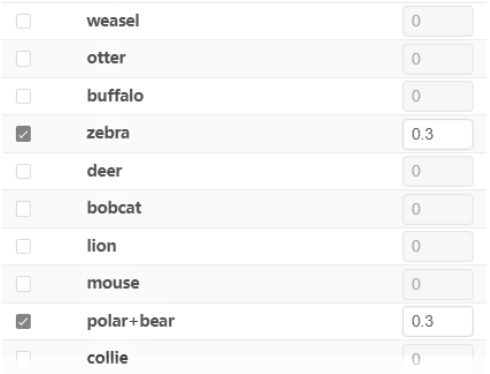


图 8 全局属性设置方式

随后，在局部编辑图选择了某一维属性，通过查看属性图像发现，该属性代表的含义似乎是黑白条纹，如图 9 所示. 本文认为 giant+panda 类在这个属性上应该具有较高的属性值，故通过拖拽的方式将其放到值较高的区间中. 在局部属性编辑图中再次选中一维值非零的属性，查看属性图像发现，该属性代表的含义应该是水面. 如图 10 所示，giant+panda 类不应该与该属性具有某种关联，因为 giant+panda 类的图像中很少会有大量水面元



图 9 黑白条纹属性展示



图 10　无关属性展示

素出现的情况. 导致 giant+panda 类该维属性不为零的原因可能是全局属性设置时选择的 polar+bear 类, 该类的图像有较大概率会有水面元素出现, 该维属性值不为零是合理的. 分析完毕后通过拖拽的方式重新设置属性值. 而在设置 humpback+whale 类的属性值时, 水面元素则具有极高的重要性, 几乎所有的该类图像都会包含水面元素, 因此在单独设置 humpback+whale 类该维属性时需要将其拖拽至高属性值区间中. 不断重复以上操作, 并完成所有未见类属性向量编辑.

使用本文系统所耗费的时间成本和人力成本更少, 并为 AWA2 数据集创建了一个由 98 个属性组成的连续值类属性矩阵, 与数据集自带的 85 个属性的二元值类属性相比, 其具有相近的训练准确度和测试准确度.

4.2　用户反馈

为了进一步验证本文方法和系统视图的有效性和实用性, 本节通过用户评估实验进行验证.

4.2.1　参与者

邀请了 20 名计算机视觉相关专业的参与者对本系统进行可行性评估, 参与者包括 3 名本科生、15 名硕士研究生和 2 名教授, 所有参与者都熟悉或理解深度学习和可视化技术.

4.2.2　评估任务

整个评估过程包括: 对于每个参与对象, 首先向参与者解释零样本图像分类的含义, 并详细介绍本系统各个视图的作用和操作流程, 讲解具体的例子; 然后为每个参与者提供一个相同的实验环境. 实验的数据集为 CUB200-2011[40], 该数据集共有 11 788 幅鸟类图像, 包含 200 类鸟类子类; 选择其中前 150 类作为已见类进行模型训练和属性提取, 实验模型为预训练的 ResNet50 模型[38], 模型的 Layer4 作为提取高维特征的中间层. 为减轻实验参与者的测试强度, 同时也为保证实验结果的可靠性, 只需要参与者通过本系统设计 10 个未见类的属性向量. 在参与者结束实验后, 需填写对本系统的调查评分问卷, 本文使用 5 等级李克特量表进行评估. 要求参与者使用 1~5 分在问卷上对 6 个问题进行评分; 5 分代表非常满意, 4 分代表满意, 3 分代表普通, 2 分代表不满意, 1 分代表非常不满意. 问题描述如下:

问题 1. 系统视图是否美观且容易理解?

问题 2. 系统视图是否清楚地显示各属性的重要性程度?

问题 3. 系统视图是否能辅助用户进行属性选择?

问题 4. 系统视图是否能辅助用户高效地编辑未见类属性?

问题 5. 系统视图是否能对用户操作给出有效反馈?

问题 6. 系统是否可以辅助设计有效的类属性矩阵?

每名参与者的实验大概持续 30 min.

4.2.3　实验结果

实验结束后, 对参与者设计的 10 个属性向量和数据集自带的属性向量, 使用同一零样本分类模型进行测试. 实验结果表明, 对于具有较高专业知识水平的用户, 本系统可以辅助其设计用于零样本图像分类的类属性矩阵. 在该实验场景下, 2 名教授参与者设计的未见类属性向量比其他参与者有更高的测试准确度; 不同参与者设计的各未见类最高测试准确度能达到数据集自带的属性向量水平.

统计参与者完成的调查问卷, 结果如图 11 所示. 大多数参与者对本文的工作表示认可, 认为通过深度学习可解释性技术提取类别属性特征是一项有价值的探索. 在零样本学习中应用这种属性提取方法最大限度地保留了属性的视觉信息, 同时也没有丧失属性的可理解性. 关于系统的易用性, 参与者也给予了高度评价, 认为系统呈现信息的方式既不会引起视觉混乱, 也没有损失关键信息, 可以协助用户快速地分析属性之间的分布趋势, 并且多个视图之间的交互性也非常强.

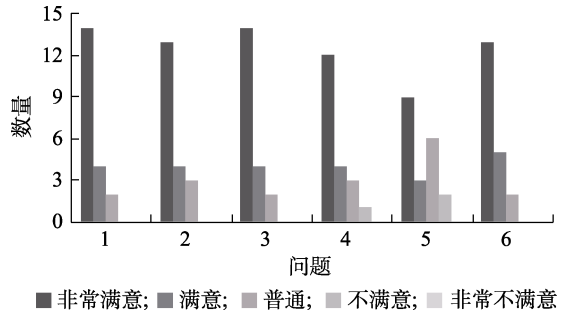


图 11 调查问卷结果

此外，参与者也对进一步优化系统提出一些建议. 系统可扩展性还存在不足，为了适应分析更大规模的数据集，系统可视化设计方面还需进一步完善. 同时，也应该重点优化系统的响应时间，以提高用户的工作效率；系统中还可以添加更多探索功能，例如，概念降维图包含的功能较少，可以添加概念集群分析的功能，使用户可以了解各个概念集群更具体信息.

5 结 语

本文以提高领域专家设计类属性矩阵的效率为出发点，首先总结分析了任务的设计需求和难点，其次提出了一种为零样本图像分类设计属性辅助信息的方法，并设计了相对应的可视化交互系统；本文方法使用基于概念的深度学习可解释性方法来获取类别属性，可以帮助领域专家高效地完成类别属性设计工作.

当然，本文的可视化方法也存在一些不足之处. 一方面，对不同数据集和不同模型的组合需要提前进行模型训练工作，虽然这不会对后续的系统使用造成影响，但仍然是一个需要优化的方向. 另一方面，对未见类属性向量的编辑操作没有一个更直接的反馈信息展示，仅仅通过属性向量间的距离不能直观地表达出每一步调整对零样本图像分类结果的影响. 未来可以考虑使用训练速度足够快的零样本图像分类模型给用户更直接的反馈，最终达到帮助使用者快速地设计类属性矩阵的目的.

参考文献(References):

[1] Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 951-958

[2] Palatucci M, Pomerleau D, Hinton G, *et al*. Zero-shot learning with semantic output codes[C] //Proceedings of the 22nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2009: 1410-1418

[3] Akata Z, Perronnin F, Harchaoui Z, *et al*. Label-embedding for image classification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1425-1438

[4] Schönfeld E, Ebrahimi S, Sinha S, *et al*. Generalized zero- and few-shot learning via aligned variational autoencoders[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 8239-8247

[5] Xian Y Q, Lampert C H, Schiele B, *et al*. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2251-2265

[6] Farhadi A, Endres I, Hoiem D, *et al*. Describing objects by their attributes[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 1778-1785

[7] Yu F X, Cao L L, Feris R S, *et al*. Designing category-level attributes for discriminative visual recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2013: 771-778

[8] Mall U, Hariharan B, Bala K. Field-guide-inspired zero-shot learning[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9526-9535

[9] Jia S C, Li Z Y, Chen N, *et al*. Towards visual explainable active learning for zero-shot classification[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 791-801

[10] Xu W J, Xian Y Q, Wang J N, *et al*. VGSE: visually-grounded semantic embeddings for zero-shot learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 9306-9315

[11] Mishra A, Krishna Reddy S, Mittal A, *et al*. A generative model for zero shot learning using conditional variational autoencoders[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2188-2196

[12] Romera-Paredes B, Torr P H S. An embarrassingly simple approach to zero-shot learning[C] //Proceedings of the 32nd International Conference on Machine Learning. Cambridge: JMLR, 2015: 2152-2161

[13] Xian Y Q, Akata Z, Sharma G, *et al*. Latent embeddings for zero-shot classification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 69-77

[14] Zhang L, Xiang T, Gong S G. Learning a deep embedding model for zero-shot learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2021-2030

[15] Liu S C, Long M S, Wang J M, *et al*. Generalized zero-shot

learning with deep calibration network[C] //Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2018: 2009-2019

[16] Kodirov E, Xiang T, Gong S G. Semantic autoencoder for zero-shot learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3147-3183

[17] Changpinyo S, Chao W L, Gong B Q, *et al*. Synthesized classifiers for zero-shot learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 5327-5336

[18] Verma V K, Rai P. A simple exponential family framework for zero-shot learning[C] //Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Heidelberg: Springer, 2017: 792-808

[19] Chen L, Zhang H W, Xiao J, *et al*. Zero-shot visual recognition using semantics-preserving adversarial embedding networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1043-1052

[20] Zhu Y Z, Elhoseiny M, Liu B C, *et al*. A generative adversarial approach for zero-shot learning from noisy texts[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1004-1013

[21] Xian Y Q, Lorenz T, Schiele B, *et al*. Feature generating networks for zero-shot learning[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5542-5551

[22] Verma V K, Arora G, Mishra A, *et al*. Generalized zero-shot learning via synthesized examples[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 4281-4289

[23] Sohn K, Yan X C, Lee H. Learning structured output representation using deep conditional generative models[C] //Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 3483-3491

[24] Bergamo A, Torresani L, Fitzgibbon A. PICODES: learning a compact code for novel-category recognition[C] //Proceedings of the 24th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2011: 2088-2096

[25] Berg T L, Berg A C, Shih J. Automatic attribute discovery and characterization from noisy web data[C] //Proceedings of the 11th European Conference on Computer Vision. Heidelberg: Springer, 2010: 663-676

[26] Parikh D, Grauman K. Interactively building a discriminative vocabulary of nameable attributes[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2011: 1681-1688

[27] Duan K, Parikh D, Crandall D, *et al*. Discovering localized attributes for fine-grained recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 3474-3481

[28] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[OL]. [2023-06-18]. https://arxiv.org/abs/2010.11929

[29] Brendel W, Bethge M. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet[OL]. [2023-06-18]. https://arxiv.org/abs/1904.00760

[30] Huang J B, Mishra A, Kwon B C, *et al*. ConceptExplainer: interactive explanation for deep neural networks from a concept perspective[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 831-841

[31] Koh P W, Nguyen T, Tang Y S, *et al*. Concept bottleneck models[C] //Proceedings of the 37th International Conference on Machine Learning. Cambridge: JMLR, 2020: Article No.495

[32] Caruana R, Lou Y, Gehrke J, *et al*. Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission[C] //Proceedings of the 21th International Conference on Knowledge Discovery and Data mining. New York: ACM Press, 2015: 1721-1730

[33] Koh P W, Liang P. Understanding black-box predictions via influence functions[C] //Proceedings of the 34th International Conference on Machine Learning. Cambridge: JMLR, 2017: 1885-1894

[34] Kim B, Wattenberg M, Gilmer J, *et al*. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C] //Proceedings of the 35th International Conference on Machine Learning. Cambridge: JMLR, 2018: 2668-2677

[35] Ghorbani A, Wexler J, Zou J, *et al*. Towards automatic concept-based explanations[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: Article No.832

[36] Ge Y H, Xiao Y, Xu Z, *et al*. A peek into the reasoning of neural networks: interpreting with structural visual concepts[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 2195-2204

[37] van der Maaten L, Hinton G. Visualizing data using *t*-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605

[38] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[39] Osherson D N, Stern J, Wilkie O, *et al*. Default probability[J]. Cognitive Science, 1991, 15(2): 251-269

[40] Wah C, Branson S, Welinder P, *et al*. The caltech-UCSD birds-200-2011 dataset[R]. Pasadena: California Institute of Technology, 2011