计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330231
60(8): 1754−1767, 2023

ADIC: 一种面向可解释图像识别的自适应解纠缠 CNN 分类器

赵小阳[1] 李仲年[1] 王文玉[1] 许新征[1,2]
[1](中国矿业大学计算机学院 江苏徐州 221116)
[2](中国矿业大学教育部矿山数字化工程研究中心 江苏徐州 221116)
(xuxinzh@163.com)

ADIC: An Adaptive Disentangled CNN Classifier for Interpretable Image Recognition

Zhao Xiaoyang[1], Li Zhongnian[1], Wang Wenyu[1], and Xu Xinzheng[1,2]
[1](*School of Computer Science and Technology*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116)
[2](*Engineering Research Center of Digital mine*, *China University of Mining and Technology*, *Ministry of Education*, *Xuzhou*, *Jiangsu* 221116)

Abstract In recent years, convolutional neural network (CNN), as a typical deep neural network model, has achieved remarkable results in computer vision fields such as image recognition, target detection and semantic segmentation. However, the end-to-end learning mode of CNNs makes the logical relationships of their hidden layers and the results of model decisions difficult to be interpreted, which limits their promotion and application. Therefore, the research of interpretable CNNs is of important significance and application value. In order to make the classifier of CNNs interpretable, many researches have emerged in recent years to introduce basis concepts into CNN architectures as plug-in components. The post-hoc concept activation vector methods take the basis concept as their representation and are used to analyze the pre-trained models. However, they rely on additional classifiers independent of the original models and the interpretation results may not match the original model logic. Furthermore, some existing concept-based ad-hoc interpretable methods are too absolute in handling concepts in the latent classification space of CNNs. In this work, a within-class concepts graphs encoder (CGE) is designed by introducing a graph convolutional network module to learn the basis concepts within a class and their latent interactions. The adaptive disentangled interpretable CNN classifier (ADIC) with adaptive disentangled latent space is proposed based on CGE by designing regularization terms that implement different degrees disentanglement of the basis concepts with different dependencies. By embedding ADIC into ResNet18 and ResNet50 architectures, classification experiments and interpretable image recognition experiments on Mini-ImageNet and Places 365 datasets have shown that ADIC can further improve the accuracy of the baseline model while ensuring that the baseline model has self-interpretability.

Key words convolutional neural network; interpretability; category basis concept; disentangle; graph convolution network

摘 要 近年来,卷积神经网络(convolutional neural network, CNN)作为一种典型的深度神经网络模型,在图像识别、目标检测和语义分割等计算机视觉领域中取得了令人瞩目的成效.然而,CNN 端到端的学习模式使其隐藏层的逻辑关系以及模型决策结果难以被解释,这限制了其推广应用.因此,研究可解释的

收稿日期: 2023-03-31; **修回日期**: 2023-06-08
基金项目: 国家自然科学基金项目(61976217); 中央高校基本科研业务费专项资金(2019XKQYMS87)
This work was supported by the National Natural Science Foundation of China (61976217) and the Fundamental Research Funds for the Central Universities(2019XKQYMS87).
通信作者: 许新征(xxzheng@cumt.edu.cn)

CNN 具有重要意义和应用价值. 为了使 CNN 的分类器具有可解释性, 近年来涌现出了很多在 CNN 架构中引入基础概念作为插入式成分的研究. 事后概念激活向量方法以基础概念为表现形式, 用于分析预训练的模型, 但依赖独立于原始模型的额外的分类器, 解释结果可能并不符合原始模型逻辑. 另外, 现有的一些基于概念的事前可解释方法对于 CNN 潜在分类空间中的概念处理太过绝对. 引入图卷积网络模块, 设计了一种类内概念图编码器（within-class concepts graphs encoder, CGE）学习类内基础概念及其潜在交互. 在 CGE 基础上, 设计实现不同依赖关系的基础概念不同程度解纠缠的正则化项, 提出了潜在空间自适应解纠缠的可解释 CNN 分类器（adaptive disentangled interpretable CNN classifier, ADIC）. 将 ADIC 嵌入 ResNet-18 和 ResNet-50 架构, 在 Mini-ImageNet 和 Places365 数据集上的分类实验和可解释图像识别实验结果表明, ADIC 在保证基准模型具有自解释能力的前提下, 可以进一步提高基准模型的精度.

关键词 卷积神经网络; 可解释性; 类别基础概念; 解纠缠; 图卷积网络

中图法分类号 TP181

近年来, 对于卷积神经网络(CNN)系列黑盒模型, 研究者们提出了越来越多的可解释方法, 其中一个主流研究方向是可视化 CNN 隐藏层中的特征表示. 然而, 神经网络的特征可视化与神经网络的语义解释之间仍存在巨大差距. 对于一个对象实例的判断, 人类通常是将该实例分解为对象部分, 并与存储在脑中的概念进行匹配, 作为识别各对象部分的证据, 用这些脑海中已识别的概念解释推理过程, 做出最终决定. 仿照人脑识别物体机制, 确定对象部分并构造概念以实现可解释的智能机器模型是一个有潜在研究价值的新兴方向. 本文中提到的“概念”在图像识别任务中的本质是“视觉概念”, 即具有语义信息且对模型预测起重要作用的像素集. 同一视觉概念在不同图像中的表现形式相似, 不同视觉概念具有不同的语义信息. 例如, 斑马的条纹、汽车的轮胎以及鸟类的羽毛等都可以作为其类别的一个基础视觉概念. 本文将视觉概念统一简称为概念.

基于概念解释模型, 一个重要问题就是如何量化定义概念. 用概念激活向量进行测试(testing with concept activation vectors, TCAV) [1] 是最早提出使用概念激活向量(concept activation vectors, CAV)量化定义概念的方法, CAV 不再分析网络单个节点的特征, 而是尝试学习它们的线性组合来表示预定义的概念. ACE[2] 基于 TCAV, 通过聚类图像块自动发现定义新概念. ICE[3] 通过对特征图进行非负的矩阵分解修改 ACE 框架, 为不同实例提供一致的 CAV 权重, 提供概念保真度测量措施. 上述 3 种基于概念的解释方法也称概念向量方法, 都针对预训练模型进行事后分析, 且都依赖于概念的潜在空间中存在一个易于分类的分类器的假设. 然而, 网络的潜在空间并无此特性, 即概念向量方法其实是基于一个独立于模型的额外的分类器. 理想情况下, 一个可解释的 CNN 不应该求助于额外的分类器, 而是具有可解释的分类器, 即可解释 CNN 的潜在分类空间如何分类概念(解纠缠)对于用户来说应该是透明的或可理解的.

现如今构造 CNN 透明潜在分类空间的代表性方法有概念白化(concept whitening, CW)[4]、TesNet[5] 和 Deformable ProtoPNet[6] 等. CW 模块通过在网络训练过程中强制约束潜在空间的轴与预定义的类别概念对齐, 约束不同的类别概念轴方向彼此正交, 从而使潜在空间中的类别概念解纠缠. TesNet 引入正交损失以鼓励类内的不同概念之间彼此正交, 在 Grassmann 流形上构造透明潜在空间. Deformable ProtoPNet 受 TesNet 启发, 在原型零件之间引入正交损失, 鼓励类内的所有原型零件彼此之间正交. CW 模块使潜在空间中的所有滤波器的输出完全去相关, TesNet 和 Deformable ProtoPNet 使同一类别内的概念彼此正交, 这些约束要求太过绝对. 在实践中有很多概念之间是高度相关或者有相对稳定的空间关系, 如“飞机”概念和“天空”概念, “车身”与“车轮”的位置关系等. 类比人脑识别物体机制, 除了对物体本身各部位的识别外, 通常还会参照物体所处环境以及参照物等信息.

因此, 在保留类别相关概念依赖关系的前提下实现类别的分离, 本文引入图卷积神经网络模块构造自适应解纠缠的透明潜在空间, 设计了一种可解释的 CNN 分类器. 采用无监督方式自动获取类别的基础概念信息, 经可解释 CNN 分类器, 自主完成不同类别及不相关概念之间的解纠缠. 本文的主要贡献有 2 点:

1)引入图卷积神经网络模块, 设计了类内概念图编码器(within-class concepts graphs encoder, CGE)自动获取类别基础概念, 以图结构形式编码类内概

念信息及概念之间的空间信息,学习类内基础概念之间的潜在交互.

2)在 CGE 编码器之后,设计了一个自适应解纠缠的可解释 CNN 分类器(adaptive disentangled interpretable CNN classifier, ADIC),通过设置三段阈值将潜在空间中的基础概念分为类内相关概念、类内不相关概念和不同类别概念 3 种类型,保留相关概念的依赖关系,在不相关概念之间添加强制性正交约束,从而实现类别概念自适应解纠缠.

1 相关工作

目前提高神经网络可解释性的研究主要分为对现有模型的事后可解释性分析(post-hoc explainability analysis)和直接构建固有事前可解释模型(ad-hoc interpretable modeling)2 个方向.

1.1 事后可解释方法

事后可解释方法通常是借助模型额外的辅助信息对训练好的神经网络模型节点激活总体趋势的统计,是为了详述黑盒模型内部功能或决策原因而采取的一些行动. 根据解释方法的解释目标是模型整体逻辑还是单个输入样本,可以将事后可解释方法分为全局解释和局部解释[7].

全局解释旨在解释模型内部的整体逻辑和工作机制[8-9],主要方法包括激活最大化、代理模型和概念激活向量方法等. 激活最大化的思路是合成最大程度激活模型整体或感兴趣神经元输出的输入模式,即表示类别特征的抽象图像. 激活最大化方法只能用于连续性数据,例如 DeepDream 算法[10]. 代理模型指的是构造一个可解释的更简单的模型模拟原始网络模型决策,包括网络压缩[11-12]、知识蒸馏[13-15]和直接提取[16-18]. 概念激活向量方法的主要思路是基于“视觉概念”,将一组具有相似特征的图像块或图像称之为一个“视觉概念”,例如一组包含条纹的图像块或图像即代表“条纹”概念. 谷歌研究团队提出的 TCAV[1]使用视觉概念进行全局解释,利用方向导数量化模型预测结果对沿着 CAV 方向变化的特定视觉概念的敏感度全局定量评估每个视觉概念对模型预测结果的影响度.TCAV 需要预先定义感兴趣的概念,通过手动收集可以表示特定概念的示例集训练线性分类器,以习得 CAV. 针对 TCAV 需要手动收集视觉概念的问题,Amirata 等人[2]提出了一种自动视觉概念提取方法 ACE,通过图像块聚类定义新概念以自动提取视觉概念,然后利用 TCAV 对提取的视觉概念进行评估. Zhang 等人[3]提出了基于可逆概念的 ICE 框架,采用非负矩阵分解可以为不同实例的相同特征提供一致的 CAV 权重,并提出一致的保真度测量措施. 局部解释通常表现为可视化解释,即以显著图或热力图的形式突出显示输入图像中对预测结果起重要作用的像素区域[19-20]. 除此之外,Liu 等人[21]还提出了稀疏对比编码(sparse contrastive coding, SCC),通过模型每一层的隐藏状态得到词向量的特征重要性,自适应地将输入分为前景和背景的任务相关性,采用监督对比学习损失提高模型可解释性和性能. 局部解释方法可以大致分为基于扰动的前向传播显著性方法[22-25]、基于反向传播的显著性方法[26-29]和基于类激活映射的显著性方法[30-33]这 3 种类型.

1.2 事前可解释方法

事前可解释方法是从头设计可以自解释的固有可解释神经网络,自解释模型在应用的同时为用户提供模型输出的决策原因,无需添加额外的信息. 事前可解释方法可以避免事后解释方法不忠实于原始模型的偏见,因为事后可解释分析中原始模型预测期间不使用事后解释,预测和解释是 2 个独立的过程. 事前可解释方法可以进一步分为模型翻新和可解释表示[34].

模型翻新是指设计模型可解释组件或新的网络结构编码特定的语义概念,实现模型内置可解释性. 例如,Chen 等人[35-36]设计基于案例推理的神经网络结构来剖析图像,通过类别典型特征解释模型推理. Wang 等人[5]提出可解释的深度模型 TesNet,构造类别子空间分离的透明潜在空间,并约束类内概念彼此正交. Jon 等人[6]提出 Deformable ProtoPNet,提供空间灵活的可变形原型,可以捕捉到目标对象的姿势变化和环境,相比 ProtoPNet[35]具有更加丰富的解释. Peng 等人[37]提出了类别可解释的神经聚类(interpretable neural clustering, TELL)网络,其将 k 均值目标重新表述为神经层,实现了算法透明化. 可解释表示通常是采用正则化技术在神经网络训练过程中学习更具可解释性的语义表示,从模型的可分解性、单调性和稀疏性等方面设计正则化项,实现模型内部表征解纠缠. 例如,Zhang 等人[38]设计了一种将每个滤波器响应约束到高层卷积层中特定对象部分的正则化损失,获取解纠缠表示. Lage 等人[39]提出新颖的 human-in-the-loop 正则化项,通过用户评估已完成训练的多个网络模型的响应时间来衡量对模型的理解程度,选择用户响应时间最短的模型. Chen 等人[4]提出一种概念白化模块,直接约束潜在空间,强制潜在空间的

轴与预定义概念对齐，不同概念之间彼此正交.

基于模型翻新技术的 TesNet 和 Deformable ProtoPNet 以及基于可解释表示技术的概念白化模块，均以视觉概念为中间形式实现模型可解释图像识别. 概念白化模块、TesNet 和 Deformable ProtoPNet 强制约束类内基础概念彼此正交甚至基础概念块间彼此正交，约束过于绝对，忽略了基础概念之间可能存在高依赖度的潜在交互. 另外，概念白化模块使用预定义概念，概念集和模型训练集相互独立. 本文针对上述问题，设计了一种自适应解纠缠的可解释分类器，通过引入图结构学习类内基础概念特征及其之间的依赖关系，对具有不同依赖度的基础概念进行不同程度的正则化约束，在保留高依赖度基础概念之间潜在交互信息的同时，实现类内不相关概念及不同类别概念的解纠缠，即实现透明化潜在分类空间.

2 基础理论

2.1 ICE 框架

基于概念的可逆解释(invertible concept-based explanations, ICE)框架是一个为预训练的 CNN 模型提供局部和全局概念级解释的框架，它采用非负矩阵分解提出非负概念激活向量(non-negative concept activation vectors, NCAV)，为特征提供一致的权重和一致的保真度测量.

ICE 框架主要由 CNN 模型分割、特征图降维器以及 CAV 权重评估这 3 部分组成，框架如图 1 所示. 首先选定预训练 CNN 的目标层l，将其分解为概念提取器E和分类器C，概念提取也就是高维特征提取. n个输入图像$\boldsymbol{I}$经特征提取得到尺寸为$n\times h\times w\times c$($h$和$w$为$A_l$的大小，$c$为通道数)的特征图$A_l$，$E_l(I)=A_l$，对特征图$A_l$采用矩阵分解进行降维，先将特征图$A_l$展平为非负矩阵$\boldsymbol{V}\in\mathbb{R}^{(n\times h\times w)\times c}$. 接着将$\boldsymbol{V}$分为特征分数$\boldsymbol{S}\in\mathbb{R}^{(n\times h\times w)\times c'}$和有意义的 NCAV $\boldsymbol{P}\in\mathbb{R}^{c'\times c}$，$\boldsymbol{V}=\boldsymbol{SP}+U$；最小化残差$U$，$\min_{S,P}\|\boldsymbol{V}-\boldsymbol{SP}\|_{\mathrm{F}}\mathrm{s.t.}\boldsymbol{S}\geqslant\boldsymbol{0},\boldsymbol{P}\geqslant\boldsymbol{0}$. 最后，建立分类器$C$的线性近似，评估每个 CAV 的重要性.

对于分类器C的特征重要性的评估，ICE 采用 TCAV[1] 中求方向导数的方法. 给定目标层l中已学习的 NCAV P_l，对于给定的特征图A_l，针对k类的权重计算如式(1)所示.

$$\frac{\partial C_{l,k}}{\partial P_l}=\frac{1}{n\times h\times w}\sum_{a\in A_l}\lim_{\varepsilon\to 0}\frac{h_{l,k}(a+\varepsilon P_l)-h_{l,k}(a-\varepsilon P_l)}{2\varepsilon}. \quad (1)$$

ICE 克服了自动概念提取算法 ACE 通过聚类特征图获取的概念权重不一致的缺点，自动获取高概念分数的类别视觉概念，但其依然需要依赖独立于原始模型的额外的分类器. 针对此局限性，本文摒弃 ICE 的权值评估部分，不使用额外的分类器，设计可自解释的神经网络模型.

2.2 图卷积神经网络

图卷积神经网络(graph convolution neural network, GCN)是 CNN 针对非欧几里德数据(也称之为图数据)衍生出的网络. 拓扑自适应图卷积网络(topology adaptive graph convolutional network, TAGCN)是 Du 等人[40] 提出的针对有向图任务进行处理的 GCN 模型，其通过设计一组固定大小(大小为 1~k)的可学习滤波器执行图上卷积，而不是对图上卷积取近似.

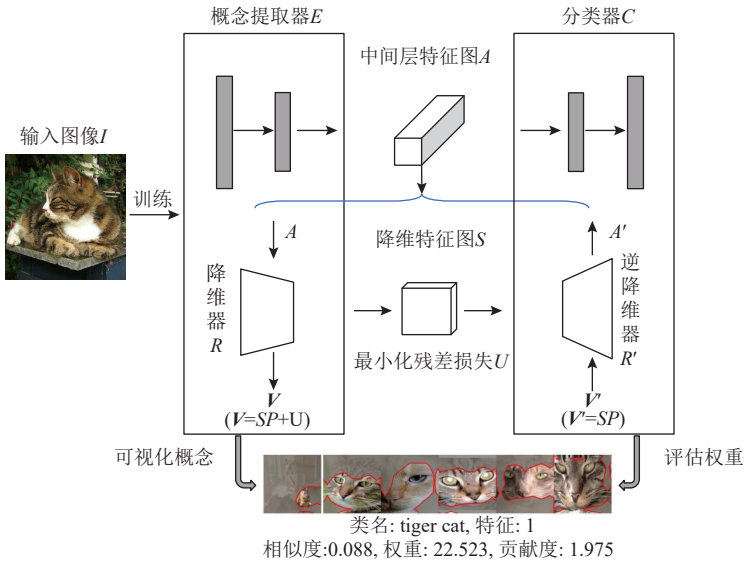

Fig. 1 ICE framework

图 1 ICE 框架

给定有向图$\mathcal{G}$上的信息及其关系表示为$\mathcal{G}=(\mathcal{V},\mathcal{E},\bar{\boldsymbol{A}})$，$\mathcal{V}$为顶点集，$\mathcal{E}$为边集，$\bar{\boldsymbol{A}}$为图的加权邻接矩阵，$\bar{A}_{n,m}$表示顶点$n$到顶点$m$的有向边权值. 第$f$个多项式第$c$个特征图卷积滤波器表示为$G_{c,f}^{(l)}=\sum_{k=0}^{K}g_{c,f,k}^{(l)}A^k$，$g_{c,f,k}^{(l)}$为滤波器多项式系数，$\mathbf{1}_{N_l}$表示数值全为1的$N_l$维向量. 输出特征图为来自不同大小滤波器的卷积结果的加权和，$\tilde{j}$是从顶点j到顶点i的所有长度为k的路径，如式(2)所示.

$$y_f^{(l)}(i)=\sum_{k=1}^{K_l}\sum_{c=1}^{C_l}\sum_{j\in\{\tilde{j}\}}(g_{c,f,k}^{(l)}\omega\left(p_{j,i}^k\right)\boldsymbol{x}_c^l(j)+b_f\mathbf{1}_{N_l}),\tag{2}$$

其中，b_f为可学习的偏差，$K_l\in\{1,2,3,\cdots\}$即图滤波器的尺寸，$\omega\left(p_{j,i}^k\right)$表示从顶点$j$到顶点$i$的所有长度为$k$的路径权重之和，$\boldsymbol{x}_f^{(l+1)}=\sigma\left(y_f^{(l)}\right)$，$\sigma(\cdot)$表示应用于顶点值的激活函数.

2.3 概念白化模块

CW模块是Chen等人[4]在2020年提出的一种直接约束潜在空间，强制潜在空间的轴与预定义的概念对齐是使潜在空间白化(去相关和归一化)的模块. CW模块作为插入模块可以替代CNN中的普通批归一化步骤，即BN层.

CW模块由白化(whitening)变换和正交(orthogonal)变换2部分组成. 白化变换主要是对数据进行去相关和标准化，如式(3)所示. 令$\boldsymbol{Z}_{d\times n}$为$n$个样本的潜在表示矩阵，其中每一列$\boldsymbol{z}_i\in\mathbb{R}^d$包含第$i$个样本的潜在特征. 对于$k$个感兴趣的概念$c_1,c_2,\cdots,c_k$，预先定义$k$个辅助数据集$X_{c_1},X_{c_2},\cdots,X_{c_k}$，$X_{c_j}$中的样本为概念$c_j$最具代表性的样本.

$$\begin{aligned}&\Psi(\boldsymbol{Z})=W\left(\boldsymbol{Z}-\mu\mathbf{1}_{n\times1}^{\mathrm{T}}\right),\mu=\frac{1}{n}\sum_{i=1}^{n}z_i,\\&\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}=\boldsymbol{\Sigma}^{-1},\boldsymbol{\Sigma}_{d\times d}=\frac{1}{n}(\boldsymbol{Z}-\mu\mathbf{1}^{\mathrm{T}})(\boldsymbol{Z}-\mu\mathbf{1}^{\mathrm{T}})^{\mathrm{T}},\end{aligned}\tag{3}$$

其中，Ψ为白化变换，μ是样本均值，$\boldsymbol{W}_{d\times d}$是白化矩阵，$\boldsymbol{\Sigma}_{d\times d}$是协方差矩阵. 白化矩阵$\boldsymbol{W}$不唯一，通过零相位分量分析(zero-phase analysis, ZCA)和Cholesky分解等多种方式计算获得.

在对潜在空间进行白化变换后，还需在潜在空间中旋转样本，以使来自概念C_j的数据在第j个轴上高度激活. 具体地，需要找到一个正交矩阵$\boldsymbol{Q}_{d\times d}$，其列$\boldsymbol{q}_j$就是第$j$个轴，即正交变换，优化目标如式(4)所示：

$$\max_{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_k}\sum_{j=1}^{k}\frac{1}{n_j}\boldsymbol{q}_j^{\mathrm{T}}\Psi\left(\boldsymbol{Z}_{c_j}\right)\mathbf{1}_{n_j\times1}\ \text{s.t.}\ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}=I_d,\tag{4}$$

其中，$\boldsymbol{Z}_{c_j}$是$\boldsymbol{X}_{c_j}$潜在表示的$d\times n_j$的矩阵. 此正交性约束优化问题通过Stiefel流形上基于梯度的方法[41]解决.

3 ADIC设计方法

3.1 类内概念图编码器

仿照人脑识别机制，用概念解释模型，首先需要的是量化定义概念. 为了使解释忠于原始模型，不借助额外的分类器对概念进行重要性度量. 本文使用ICE提取类别概念，但不评估概念分数，而是对类别内概念进行重新聚类编码，获取类别基础概念. 接着，以有向无环带权图表示原始图像，基础概念为顶点，概念中心点间的连接为有向边，概念间的依赖度为有向边的权值. 使用TAGCN学习类内基础概念的潜在交互，设计基于GCN的类内概念图编码器CGE. CGE用不同的概念成分或不同的概念交互解释类别差异. CGE的流程如图2所示.

首先，采用无监督的K均值聚类算法，对基于ICE生成的概念样本按类别进行重新聚类编码，获取类别基础概念$B=\left\{b_j^c\right\}_{j=1,c=1}^{m\times C_n}$，$m$为每一类的基础概念数，$C_n$为类别数. 不进行基础概念的影响度评估，如图2上图所示. 接着，构造类内概念图(within-class concepts graphs, WCG)，将原始输入图像表示为有向无环带权图$\mathcal{G}=(\mathcal{V},\mathcal{E},\bar{\boldsymbol{A}})$的形式，如图2左下图所示. 以基础概念为顶点$b_j\in\mathcal{V}$；概念$b_j$中心点到概念$b_i$中心点的连接表示为有向边$\langle b_j,b_i\rangle\in\mathcal{E}$；$\bar{\boldsymbol{A}}$为加权邻接矩阵，$\bar{\boldsymbol{a}}_{j,i}\in\bar{\boldsymbol{A}}$表示顶点$j$到顶点$i$的有向边权值. 最后，添加TAGCN模块，以WCG集为输入. 基于TAGCN学习类内概念间的潜在交互，即基础概念之间的依赖度$\bar{\boldsymbol{a}}_{j,i}$，如图2右下图所示. 参照式(2)，因为WCG集仅关心一阶路径的邻居顶点，故图滤波器的尺寸仅取1. 因此，对于一个WCG，其第l层GCN层的第f个输出特征图的表示为：

$$y_f^{(l)}(i)=\sum_{c=b_1}^{b_m}\sum_{j\in N(i)}g_{c,f}^{(l)}\omega\left(\bar{\boldsymbol{a}}_{j,i}\right)\boldsymbol{x}_f^l(j)+b_f\mathbf{1}_{N_l},\tag{5}$$

其中，i表示第i个顶点，$g_{c,f}^{(l)}$为滤波器多项式系数. 向量$\boldsymbol{x}_f^l\in\mathbb{R}^{N_l}$指第$f$个特征($f=b_1,b_2,\cdots,b_m$，基础概念$b_i$)的所有顶点上的第$l$层的输入数据. N_l为第l层的顶点数. $N(i)$为顶点i的一阶相邻顶点集. $\bar{\boldsymbol{a}}_{j,i}$揭示顶点$i$和$j$之间的依赖关系，为可训练的标量可以取任意实数值或复数值. $\bar{\boldsymbol{a}}_{j,i}^c$表示对于类$c$，概念$b_i$和$b_j$之间的依赖关系. 用$\omega\left(\bar{\boldsymbol{a}}_{j,i}\right)$计算顶点$i$与领域顶点$j$的权重之和.

本文在CGE中采用2层GCN结构，即TAGCN模块含有2层隐含层. 依照TAGCN，每个GCN层包含16个图卷积滤波器，以提取图数据特征并捕获顶

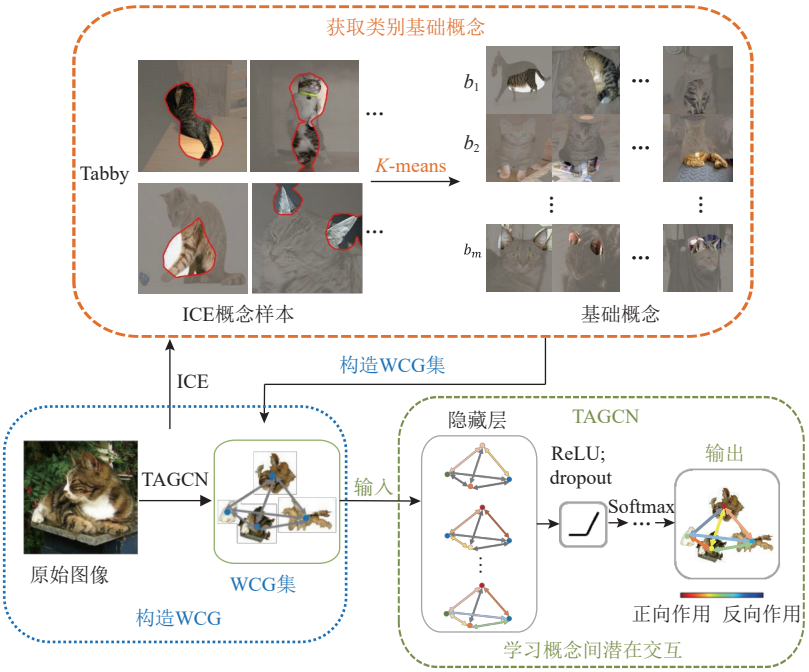

Fig. 2 CGE flow chart

图 2 CGE 流程图

点间的聚合权值. 在隐藏层之后添加一个 ReLU 激活，以进行非线性激活操作，使 GCN 层输入特征图的所有分量都为非负的. 在 ReLU 激活之后添加 dropout 操作防止过拟合. 在第 2 个 GCN 的 dropout 之后，使用 Softmax 函数获取顶点（概念）的逻辑回归值和概念间的相互依赖关系.

3.2 基于 CGE 的相关概念依赖性度量

给定一小批量输入图像集$\{x_1, x_2, \cdots, x_t\} \in X$，$\{y_1, y_2, \cdots, y_t\} \in Y$为它们的标签. X经过 CNN 特征提取子网络进行特征提取，随后经过 ICE 模块获取类别概念样本，实现类别预分离. 再经过 CGE 对类别概念样本进行重新聚类编码，保留各基础概念簇的预分离类别信息. 本文实验中在每个预分离的类别内随机选择 4 个基础概念构造一个 WCG，将输入图像集 X 经过 CGE 编码器，转变为 WCG 集$\{G_1, G_2, \cdots, G_t\} \in \mathcal{G}$.

WCG 为一个有向无环带权图$G_k = (\mathcal{V}, \mathcal{E}, \bar{\boldsymbol{A}})$. 顶点$b_i \in \mathcal{V}$为基础概念；有向边$\langle b_i, b_j \rangle \in \mathcal{E}$为 2 个基础概念中心点间的连接，其包含中心点的相对位置信息；$\bar{a}_{i,j} = \begin{cases} w_{i,j}, \langle b_i, b_j \rangle \text{ or } \langle b_j, b_i \rangle \in \varepsilon, \\ 0 \end{cases}$，$\bar{\boldsymbol{a}}_{j,i} \in \bar{\boldsymbol{A}}$中数值表示顶点$j$到顶点$i$的有向边权值.

按照预分离的类别信息，逐类别输入 GCN 模块. 使用 TAGCN 学习顶点间的聚合权值，对于每一小批量的 WCG 集 $\{G_1, G_2, \cdots, G_t\} \in \mathcal{G}$，进行式（5）的图卷积操作之后再使用一个非线性操作单元，如式（6）所示.

$$\boldsymbol{x}_f^{(l+1)} = \sigma\left(y_f^l\right), \tag{6}$$

其中，$x_f^{(l+1)}$表示第$l+1$层中含有第f个特征（特定类别下的一个基础概念$b_i \in B$）的所有顶点上的输入信息. $B = \left\{b_j^c\right\}_{j=1,c=1}^{m \times C_n}$，$m$为每一类的基础概念数，$C_n$为类别数. y_f^l表示第l层的第f个图滤波器的输出. $\sigma(\cdot)$为非线性操作单元，采用修正线性单元 ReLU 函数.

为了更好地获取基础概念之间的依赖关系，参照 Christopher 等人[42]捕捉概念间关系的思想，将 WCG 的顶点特征和边特征进行连接（concatenate）训练. 基于式（5）的顶点特征，顶点特征和边特征的连接表示如式（7）所示.

$$\begin{aligned} y_f^{(l)}(i) &= \sum_{c=b_1}^{b_m} \sum_{j \in N(i)} g_{c,f}^{(l)} \omega_1 \boldsymbol{\Gamma}\left(\omega\left(\bar{a}_{j,i}\right) x_f^l(j) + e_{ji}^l\right), \\ e_{ji}^{l+1} &= \omega_2 e_{ji}^l, \end{aligned} \tag{7}$$

其中，$\boldsymbol{\Gamma}(\cdot)$表示张量连接操作，$e_{ji}^l$表示第$l$层 GCN 中的顶点$i$和$j$的边特征，$\omega_1$和$\omega_2$分别为连接特征（顶点特征和边特征的连接特征）和边特征的线性变换参数. 通过连接训练，由整体的训练目标学习更新标量$\bar{a}_{j,i}$，衡量顶点i和j之间的依赖关系，进而获取基础概念之间的依赖度.

3.3 自适应解纠缠的可解释 CNN 分类器

由 CGE 自动获取类别基础概念，并学习初步预分离类别内基础概念之间的依赖关系. 设置概念间的依赖度阈值，将潜在空间中的类别基础概念划分

为相关概念和不相关概念. 保留相关概念的依赖关系, 通过在不同类别概念和类内不相关概念之间添加不同的解纠缠约束, 设计了具有自适应解纠缠潜在空间的可解释分类器 ADIC, 使不相关概念彼此正交、不同类别彼此分离, 即实现潜在分类空间透明化. 图 3 展示了基于 ADIC 的 CNN 框架.

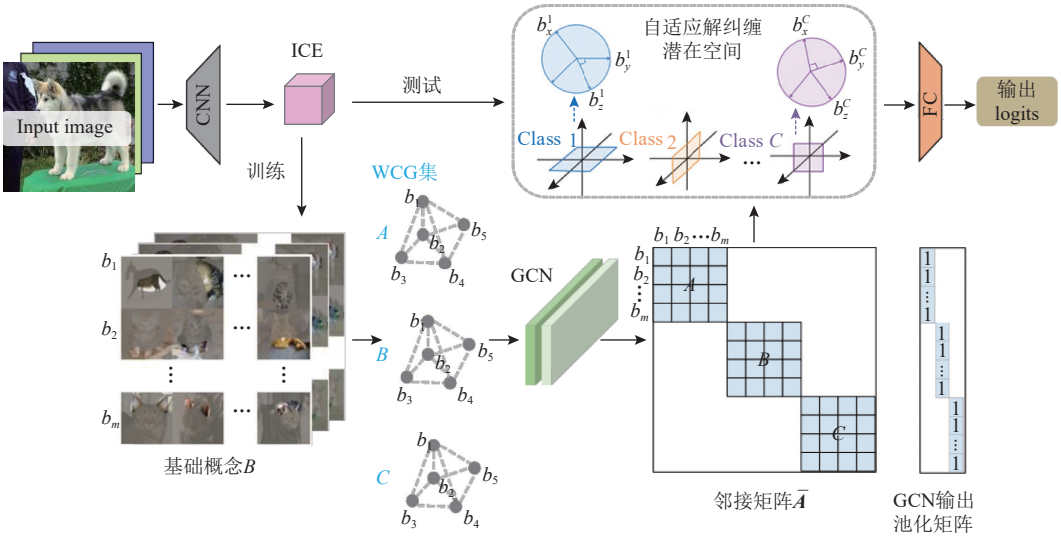

Fig. 3 ADIC-based CNN framework

图 3 基于 ADIC 的 CNN 框架

图 3 所示的基于 ADIC 的 CNN 框架具体操作流程为:

1)设置类内基础概念的依赖度阈值. 假定一小批量输入图像集$\{x_1, x_2, \cdots, x_t\} \in X$, 输入 CGE 构造对应的 WCG 集$\{G_1, G_2, \cdots, G_t\} \in \mathcal{G}$, 经 GCN 模块后将 WCG 集的邻接矩阵连接成一个稀疏的块对角矩阵, 批量处理图像集. 块对角矩阵的每一块对应一个 WCG 的邻接矩阵. 对稀疏矩阵按图级输出进行池化, 以池化矩阵形式表示; 再通过 GCN 模块最后的 Softmax 函数获取基础概念(顶点特征)以及基础概念间依赖度(边特征)的逻辑回归值(logits), 取值范围为$[-\infty, +\infty]$. 根据经验设置 logits 阈值, 将输入图像集的所有基础概念分为 3 种情况:

$$\text{if node}_i \geqslant 0.5 \begin{cases} edge_{ij} \geqslant 0.3 \text{或} edge_{ji} \geqslant 0.3, & (8(a)) \\ edge_{ij} < 0.3 \text{或} edge_{ji} < 0.3, & (8(b)) \end{cases}$$

$$\text{else node}_i < 0.5, \quad (8(c))$$

其中, $node_i$表示顶点i的 logit, $edge_{ij}$和$edge_{ji}$分别表示顶点i到顶点j的有向边 logit 和顶点j到顶点i的有向边 logit. 对于每个预分离类别的 WCG 集中的所有基础概念, 若情况是式 8(a)将被视为同类别的相关概念; 若情况是式 8(b)将被视为同类别的不相关概念; 否则将被视为不同类别的概念, 即式 8(c).

2)实现式 8(c)不同类别的基础概念正交分离. 不同类别, 即图 3 中的 Class 1、Class 2 和 Class C 参照 CW 模块通过在 Stiefel 上进行曲线搜索, 强制约束潜在空间中预定义概念彼此正交的思想. Stiefel 流形是一种特殊的黎曼流形, 由正交矩阵$\{\boldsymbol{O} \in \mathbb{R}^{n\times p} : \boldsymbol{O}^{\mathrm{T}}\boldsymbol{O} = \boldsymbol{D}\}$组成, $\boldsymbol{D}$为单位矩阵. 本文同样通过在潜在空间中寻找一个正交矩阵$\{\boldsymbol{Q} \in \mathbb{R}^{m\times m} : \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{D}_m\}$, 使不同类别的基础概念沿$\boldsymbol{Q}$的不同列方向高度激活, 采用在 Stiefel 流形上计算梯度的方法优化正交约束, 进而促使不同类别的基础概念相互正交分离.

对于一批属于情况式 8(c)的基础概念集$\{B_1, B_2, \cdots, B_t\} \in \mathcal{B}^{(c)}$, 不同类别基础概念正交约束损失$\mathcal{L}_{\text{orth}}$, 其具体数学表达式如式(9)所示.

$$\mathcal{L}_{\text{orth}} = \max_{\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_t} \sum_{k=1}^{t} \frac{1}{n_k} \sum_{b_j^k \in B_k} \boldsymbol{q}_k^{\mathrm{T}} \Psi\left(\Phi\left(b_j^k; \theta\right); \boldsymbol{W}, \mu\right),$$

$$\text{s.t. } \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{D}_m, \quad (9)$$

其中, $\boldsymbol{q}_k$为正交矩阵$\boldsymbol{Q}$的第k列. B_k表示感兴趣的k类的基础概念集, b_j^k为k类的基础概念样本. $\Phi(\cdot)$表示特征提取器, 参数为θ. $\Phi\left(b_j^k; \theta\right)$得到概念样本$b_j^k$的潜在特征$\boldsymbol{z}_k \in \mathbb{R}^m$. $\boldsymbol{Z} \in \mathbb{R}^{m\times n_k}$表示$n_k$个基础概念样本的潜在表示矩阵, $\boldsymbol{z}_k$为$\boldsymbol{Z}$的列元素. 类比 CW 模块, $\Psi(\cdot)$为白化变换, 具体表达形式如式(3)所示.

对于 Stiefel 流形上的参数矩阵优化, 通常采用 Cayley 变换交替更新. 本文采用 Cayley 变换[41]更新正交矩阵$\boldsymbol{Q}$, 具体数学表达式为:

$$\boldsymbol{Q} \leftarrow \left(\boldsymbol{D} - \frac{\alpha\left(G\boldsymbol{Q}^{\mathrm{T}} - \boldsymbol{Q}\boldsymbol{G}^{\mathrm{T}}\right)}{2}\right)^{-1}\left(\boldsymbol{D} + \frac{\alpha(G\boldsymbol{Q}^{\mathrm{T}} - \boldsymbol{Q}\boldsymbol{G}^{\mathrm{T}})}{2}\right)\boldsymbol{Q}, \quad (10)$$

其中，α为学习率，$\boldsymbol{G}$为网络分类损失函数的梯度.

3)实现式 8(b)同类别的不相关概念正交归一化分离. 对于 Class C 的不相关概念，即图 3 中的b_x^C、b_y^C和b_z^C. 对于一批属于情况式 8(b)的基础概念样本集$\left\{b_1^C, b_2^C, \cdots, b_t^C\right\} \in \boldsymbol{B}_{(b)}$，以矩阵形式表示为$\boldsymbol{B}_{(b)} \in \mathbb{R}^{t\times d}$，$b_i^C$为每一行为同一类别的一个不相关的基础概念，$d$为基础概念特征向量的维度. 本文采用正交归一化损失，使类内不相关的概念之间彼此推开.

同类别不相关基础概念正交归一化损失$\mathcal{L}_{\text{orth_norm}}$，其具体数学表达式如式(11)所示.

$$\mathcal{L}_{\text{orth\_norm}} = \sum_{c=1}^{C_b} \left\| \boldsymbol{B}_{(b)} \boldsymbol{B}_{(b)}^{\mathrm{T}} - \boldsymbol{D}_t \right\|_{\mathrm{F}}^2, \tag{11}$$

其中，C_b为存在符合情况式 8(b)基础概念的类别数. $\boldsymbol{D}_t \in \mathbb{R}^{t\times t}$为单位矩阵. $\|\cdot\|_{\mathrm{F}}$为弗罗贝尼乌斯范数(Frobenius norm)，即对矩阵内元素求平方和再开方. $\|\cdot\|_{\mathrm{F}}^2$则表示求矩阵内元素的平方和. 通过最小化$\mathcal{L}_{\text{orth_norm}}$，实现不相关概念之间的分离.

4)实现基于基础概念的分类. 在不同类别及不相关概念解纠缠分离之后，优化分类器的总体识别损失，以确保 ADIC 的分类准确性. 本文以基础概念为单位，采用标准交叉熵损失实现最终分类. 给定训练集$\{(x_i, y_i)\}_{i=1}^n$，ADIC 识别损失$\mathcal{L}_{\text{re}}$，其具体数学表达式如式(12)所示.

$$\mathcal{L}_{\text{re}} = \min_{\theta,\omega} \frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C_n} y_{ic} \log g_c(\Phi(x_i;\theta), \omega, B_c), \tag{12}$$

其中，C_n为总类别数. y_{ic}表示输入样本x_i的 one-hot 编码标签的第c个元素. $g(\cdot)$表示分类器，参数为ω，最后一层满足归一化条件$\sum_{c=1}^{C_n} g_c(\Phi(x_i;\theta), \omega, B_c) = 1$. B_c为经 CGE 编码器编码后预分离的第c类的基础概念集.

综上，ADIC 嵌入相关基础 CNN 架构进行端到端训练时，联合优化目标可表示为式(13)的形式.

$$\mathcal{L} = \mathcal{L}_{\text{re}} + \lambda_1 \mathcal{L}_{\text{orth}} + \lambda_2 \mathcal{L}_{\text{orth\_norm}}, \tag{13}$$

其中，λ_1和λ_2为网络训练过程中平衡各项的超参数.

训练完成后，实现分类空间的解纠缠，得到不同类别基础概念分离，以及同类别不相关概念分离的透明分类潜在空间. 测试时，将测试图像中的潜在图像块与解纠缠之后的各类别基础概念依据相似度进行匹配，得到属于各类别的概念相似度分数；再判断潜在图像块之间的交互关系(WCG 集的邻接矩阵，即边特征)和潜在空间中的基础概念之间的依赖关系是否相符，得到概念间关系的相似度分数；最后，将概念相似度分数和概念间关系的相似度分数的加权和作为最终的相似性度量，以此判断测试图像所属类别.

4 实　　验

为了验证本文提出的自适应解纠缠分类器 ADIC 的有效性和可解释性，本文以 VGG-16、ResNet-18 和 ResNet-50 模型为基础 CNN 架构，搭载 ADIC 分类器，在 Mini-ImageNet 和 Places365 数据集上进行实验. 分析搭载 ADIC 分类器的模型的性能表现，针对 Mini-ImageNet 特定测试实例实现可解释图像识别.

4.1 实验设置

Mini-ImageNet 数据集为 ImageNet 的部分节选，共有 100 个类，每类 600 张 RGB 图像，常用于模型设计或者小样本学习研究，满足本文验证可解释分类器的需求. Places365 数据集是一种遵循人类视觉认知原则的场景分类数据集，常用于对象识别、事物预测，以及理论推理等高级视觉理解任务. Places365 数据集共包含 365 个独特场景类别，每类 5 000~30 000 张 RGB 图像.

本文实验环境具体为：CPU 为 Intel Xeon Gold 6148，实际内存 63 GB，GPU 为 NVIDIA Tesla V100，显存 16 GB. 所有实验均采用 PyTorch 深度学习框架，使用 CUDA 10.1.所有模型都从头开始训练，均采用动量为 0.9 的随机梯度下降算法对网络模型进行优化，权值衰减率设为 0.000 1，输入批尺寸设为 64，epoch 设为 100，初始学习率设为 0.05.

4.2 基于 ADIC 分类器的 CNN 模型实验

本节验证 ADIC 分类器的解纠缠能力，将 ADIC 分别嵌入 VGG-16、ResNet-18 以及 ResNet-50 这 3 种经典 CNN 模型中. 在 Mini-ImageNet 数据集上训练 6 个模型：VGG16、ResNet18 和 ResNet50，以及添加了 ADIC 的 ADIC-VGG16、ADIC-ResNet18 和 ADIC-ResNet50. 由于本文主要采用正则化技术实现模型内部基础概念解纠缠，属于事前可解释表示方法. 因此，与同样采用可解释表示技术的 CW 模块进行对比，在 Places365 数据集上训练 6 个模型：ResNet18 和 ResNet50，添加了 CW 模块的 CW-ResNet18 和 CW-ResNet50，以及添加了 ADIC 的 ADIC-ResNet18 和 ADIC-ResNet50. 结果分别如表 1 和表 2 所示，通过 2 组对比实验，验证 ADIC 分类器的解纠缠能力，即分类能力.

从表 1 可以看出，添加 ADIC 分类器可以提高原始模型的分类精度. 相较于原始 CNN 模型，在 Top-1 正确率上，精度提高了大约 3 个百分点；在 Top-5 正

Table 1 Comparison Results on Mini-ImageNet Dataset

表 1 在 Mini-ImageNet 数据集上的对比结果 %

模型	Top-1 正确率	Top-5 正确率
VGG16	70.436	89.513
ADIC-VGG16	73.273	91.158
ResNet18	71.538	90.462
ADIC-ResNet18	74.470	92.043
ResNet50	73.017	90.718
ADIC-ResNet50	76.197	92.111

Table 2 Comparison Results on Places365 Dataset

表 2 在 Places365 数据集上的对比结果 %

模型	Top-1 正确率	Top-5 正确率
ResNet18	54.5	84.6
CW-ResNet18	53.9	84.2
ADIC-ResNet18	55.7	85.7
ResNet50	54.7	85.1
CW-ResNet50	54.9	85.2
ADIC-ResNet50	55.8	86.3

确率上，精度提高了大约 1.5 个百分点. 不同的网络结构结果具有一定差异.

从表 2 的结果可以看出，添加了 CW 模块的可解释 CNN 模型，即 CW-ResNet18 和 CW-ResNet50，它们的精度与原始模型保持 1% 的差异. 而添加本文设计的 ADIC 的 CNN 模型，ADIC-ResNet18 和 ADIC-ResNet50，精度均高于原始模型和 CW 可解释模型. 因此，得出 ADIC 分类器可以有效实现类别分离，提高模型性能.

另外，为了验证 ADIC 添加到网络不同深度层可能产生的差异，表 3 和表 4 分别展现在 ResNet18 和 ResNet50 的不同位置添加 ADIC 时模型性能的变化. 其中，该数值表示 ADIC 添加到该数值的构建块之后，如 2 即 ADIC 添加到第 2 个构建块之后.

由表 3 和表 4 可以看到，将 ADIC 添加到 ResNet 更深层的性能表现优于将其添加到 ResNet 更浅层，精度随添加的深度增加而提升. 因为层次越深，ADIC 获取到的特征越丰富. 因此，为了获得更好的模型表现，添加 ADIC 的最佳位置应选择在全连接层之前的最后一个卷积层后.

4.3 Mini-ImageNet 相似实例的可解释识别

本节以添加了 ADIC 的 ADIC-ResNet18 为主干网络，针对 Mini-ImageNet 数据集中的相同物种以及具有相似特征的不同物种的特定测试图像，通过可视化测试图像中与对比类别相关的潜在部位(类别基础概念)以及部位之间的位置交互(基础概念之间

Table 3 Results of the ADIC Located in Different Depth Layers of ResNet18

表 3 ADIC 位于 ResNet18 不同深度层的结果 %

构建块	Top-1 正确率	Top-5 正确率
2	71.034	90.179
4	72.487	90.709
6	73.578	91.288
8	74.470	92.043

Table 4 Results of the ADIC Located in Different Depth Layers of ResNet50

表 4 ADIC 位于 ResNet50 不同深度层的结果 %

构建块	Top-1 正确率	Top-5 正确率
3	75.444	92.068
8	75.752	92.060
16	76.197	92.111

的依赖关系)，实现可解释的图像识别. 对于相同物种，以 Mini-ImageNet 数据集中的 Japanese_spaniel (n02085782)、Blenheim_spaniel (n02086646)以及 Shih-Tzu (n02086240)这 3 类狗类样本为例. 图 4~6 分别展示了这 3 类样本中的一张测试图像的可解释识别过程，分别可视化与指定类别相关度前 4 的潜在图像块(顶点根据潜在图像块相关度排名指定颜色)以及相关度前 6 的位置关系(有向边颜色取决于顶点间依赖值大小).

从图 4 可以看到，对于一张真实类别为 Japanese_spaniel 的测试图像，其潜在图像块与 Blenheim_spaniel 类和 Shih-Tzu 类的相关度均高于其对于 Japanese_spaniel 类的相关度值，且其判断真实类别的前 4 个图像块中包含天空和衣服这种和所判断的真实类别明显无关的图像块. 但模型依然可以判断天空与狗之间，以及人类与狗之间的潜在交互关系(位置关系)与 ADIC 已学到的相关概念之间的依赖关系相符(有向边大多趋向于正向作用)，而与学习到的针对 Blenheim_spaniel 和 Shih-Tzu 这 2 个类别内的相关概念的依赖关系不相符(有向边大多趋向于反向作用)，因此模型最终可以正确预测该测试图像属于 Japanese_spaniel 类.

由图 4~6 的可视化结果可以得出，对于一张图像的识别，模型不仅关注感兴趣类别的基础概念(潜在图像块)，还关注基础概念之间的潜在交互(位置关系). 需要指出的是，由于 Mini-ImageNet 数据集中部分类别的训练图像包含较复杂场景，或者目标对象占整体图像区域的比例较小(例如图 4 中 Japanese_spaniel 在整张测试图像中的占比较小)，可能会导致模型最终学习到的类别基础概念并不全部来自于目

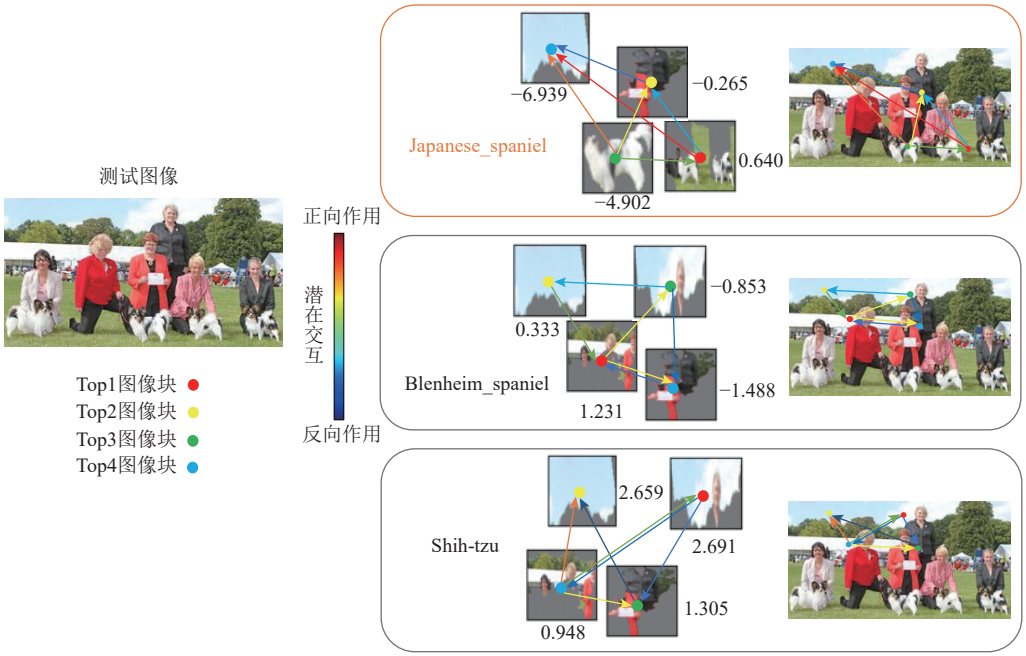

Fig. 4 Interpretable image recognition of the Japanese_spaniel class test image

图 4 Japanese_spaniel 类测试图像可解释图像识别

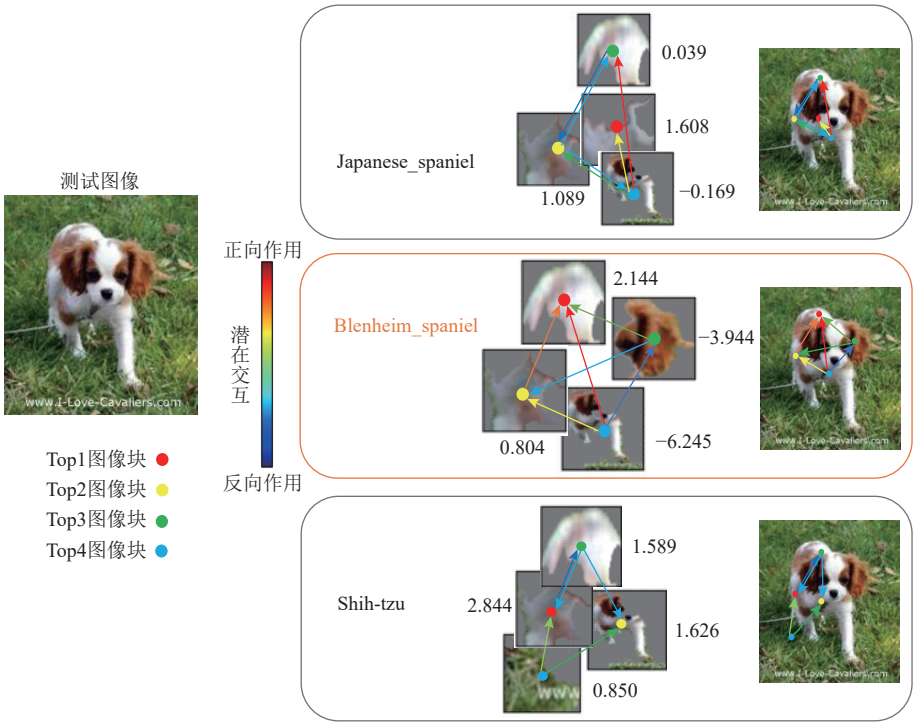

Fig. 5 Interpretable image recognition of the Blenheim_spaniel class test image

图 5 Blenheim_spaniel 类测试图像可解释图像识别

标对象本身. 尽管如此, ADIC 依旧可以准确学习到基础概念之间的位置关系, 最终通过概念相似度和概念间的相对位置关系对图像所属类别进行最终决策.

对于具有相似特征的不同物种, 以 Mini-ImageNet 数据集中的 Malamute (n02110063)和 Timber_wolf (n02114367) 以及 Tabby (n02123045)和 Snow_leopard (n02128757)这 4 张测试图像为例, 4 类 2 组样本的可解释图像识别可视化对比结果如图 7 和图 8 所示.

从图 7~8 中可以看到, 具有相似特征的不同物种对象, 例如 Malamute 和 Timber_wolf 皮毛的颜色及分布

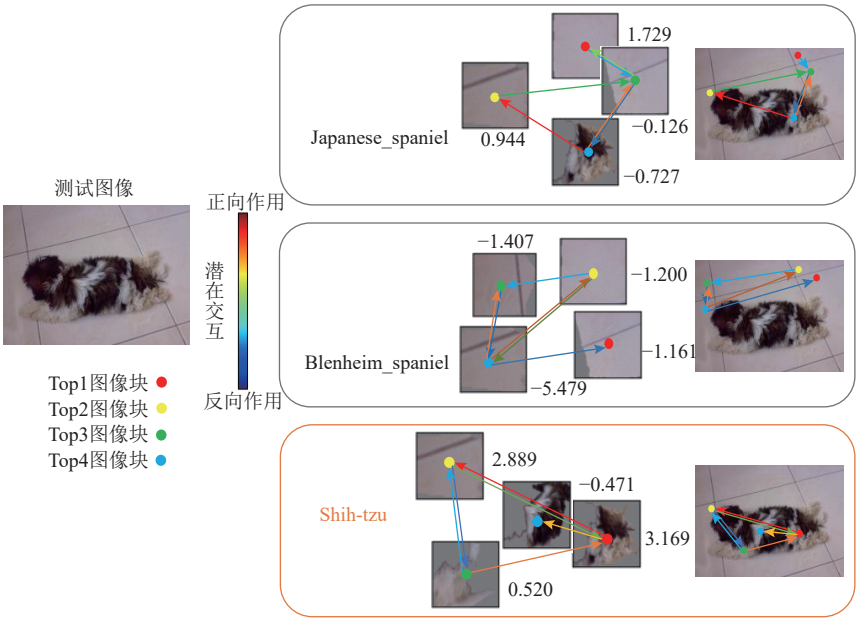

Fig. 6 Interpretable image recognition of the Shih-Tzu class test image

图 6 Shih-Tzu 类测试图像可解释图像识别

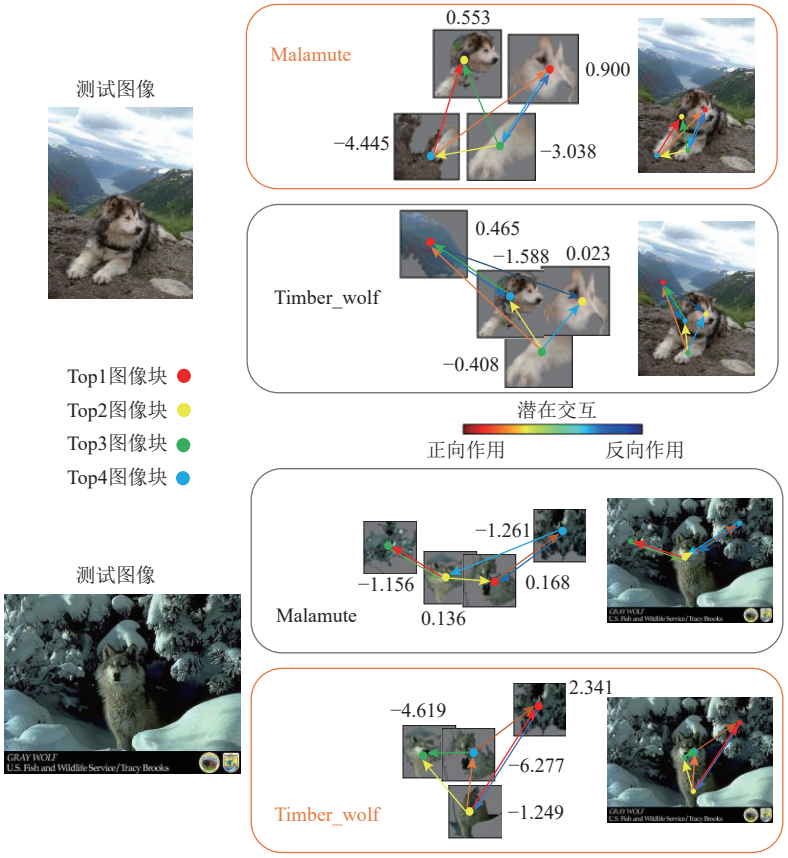

Fig. 7 Interpretable image recognition of the Malamute class and the Timber_wolf class test images

图 7 Malamute 类和 Timber_wolf 类测试图像可解释图像识别

接近，Tabby 和 Snow_leopard 具有相似的猫科动物特征(花纹、胡须等). 与相同物种的不同类别对象识别类似，模型匹配感兴趣类别的基础概念和基础概念之间的潜在交互，并做出最终预测. 对于正确预测，模型能正确聚焦到测试图像中类别对象的关键部位(类别内相关概念)，而对于其它错误类别，模型检测

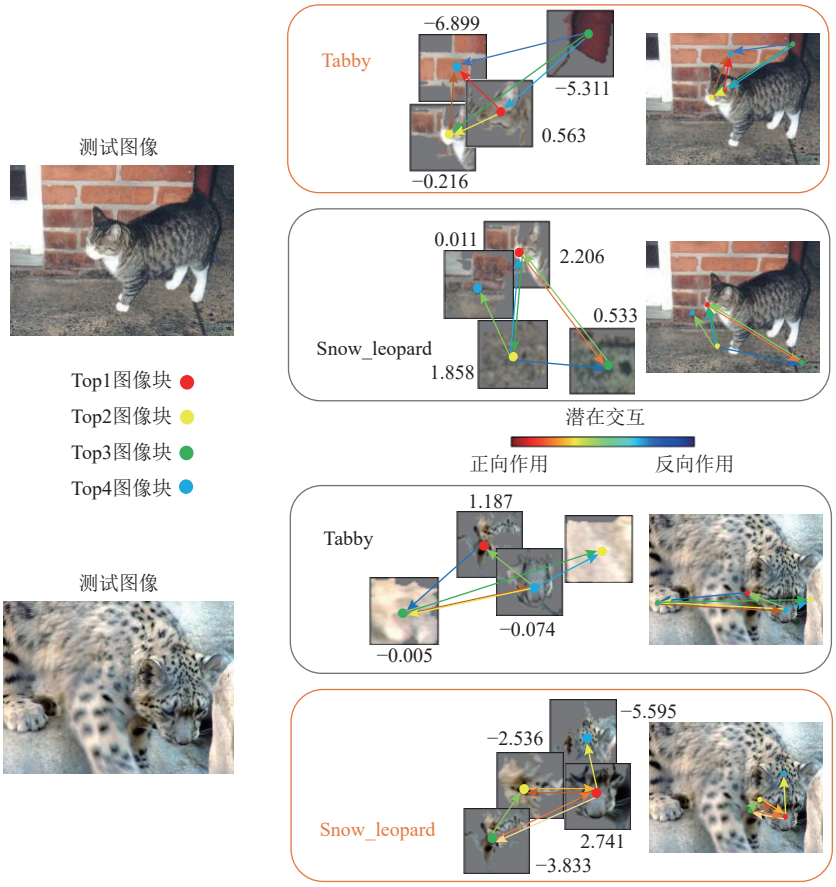

Fig. 8 Interpretable image recognition of the Tabby class and the Snow_leopard class test images

图 8 Tabby 类和 Snow_leopard 类测试图像可解释图像识别

到的大多是背景信息或者其它类别信息(不相关概念). 上述可视化图都是模型正确预测后的可视化结果, 图 9 展示了模型对于一张 Snow_leopard 类测试图像错误预测的可视化结果.

在图 9 中, 模型将真实类别为 Snow_leopard 的测试对象错误识别为 Tabby, 可以看出模型提取到的该测试图像中的潜在图像块与 Tabby 类基础概念相关度高于 Snow_leopard, 并且潜在图像块之间一半以上的位置关系(正向作用)符合已学习的 Tabby 类相关概念间的依赖关系, 因此模型做出了错误的判断.

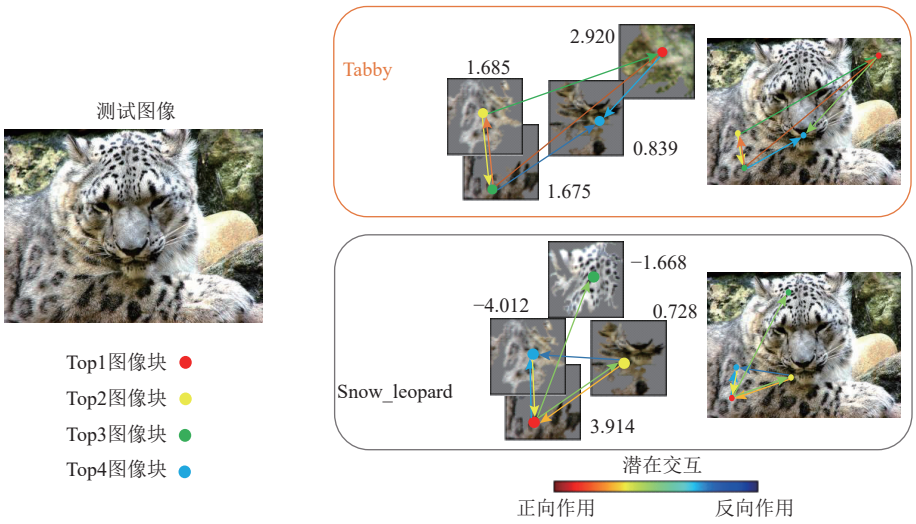

Fig. 9 Visualization result of model error identification

图 9 模型错误识别的可视化结果

综上,本节通过可视化模型决策依据(类别基础概念及概念间依赖关系),实现模型推理透明化,验证了模型的自解释性,即 ADIC 的自解释性.综合 3.2 节在 Mini-ImageNet 数据集上图像分类的实验结果,验证了添加 ADIC 分类器能进一步提高原始 CNN 模型精度,且保证了模型的自解释能力.

5 结 论

本文引入图结构设计潜在空间自适应解纠缠的可解释 CNN 分类器,在保留具有高依赖度的相关概念潜在交互的前提下,实现不相关概念的自动化分离.首先,利用 K 均值聚类算法自动获取初步预分离的各类别的基础概念.接着,引入图卷积模块设计类内概念图编码器,用有向无环带权图形式编码潜在分类空间中的特征图,获取类内基础概念的顶点特征及基础概念间的依赖关系.然后,提出了自适应解纠缠的可解释分类器 ADIC,设置三段阈值将所有基础概念样本划分为同类别相关概念、同类别不相关概念以及不同类别概念 3 部分,通过添加在 Stiefel 流形上的正交矩阵优化和正交归一化损失,依次实现不同类别基础概念分离和类内不相关概念分离.最后,将 ADIC 分别嵌入 VGG16、ResNet18 和 ResNet50 这 3 种经典 CNN 架构,在 Mini-ImageNet 数据集和 Places365 数据集上进行图像分类实验和可解释图像识别实验,实验验证了 ADIC 的适用性和可解释性,且具有较好的解纠缠能力.

作者贡献声明:赵小阳提出主要研究思路,完成实验并撰写论文;李仲年提出指导意见,参与论文修订;王文玉协助完成部分实验,参与论文修订;许新征指导论文写作,修改和审核论文.

参 考 文 献

[1] Been K, Martin W, Justin G, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)[C] //Proc of the 35th Int Conf on Machine Learning (ICML), Stockholm, Sweden: PMLR, 2018: 2668−2677

[2] Amirata G, James W, James Y Z, et al. Towards automatic concept-based explanations[C] //Proc of the Conf on Advances in Neural Information Processing Systems (NeurIPS). Cambridge, MA: MIT, 2019: 9273–9282

[3] Zhang Ruihan, Prashan M, Tim M, et al. Invertible concept-based explanations for CNN models with non-negative concept activation vectors[C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2021: 11682–11690

[4] Chen Zhi, Bei Yijie, Cynthia R. Concept whitening for interpretable image recognition[J]. Nature Machine Intelligence, 2020, 2(12): 772−782

[5] Wang Jiaqi, Liu Huafeng, Wang Xinyue, et al. Interpretable image recognition by constructing transparent embedding space[C] //Proc of the IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 875−884

[6] Jon D, Alina J B, Chen Chaofan. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes[C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 10255−10265

[7] Francesco B, Fosca G, Riccardo G, et al. Benchmarking and survey of explanation methods for black box models[J]. arXiv preprint, arXiv: 2102.13076, 2021

[8] Ji Shouling, Li Jinfeng, Du Tianyu, et al. A survey of interpretability methods, applications and security of machine learning models[J]. Journal of Computer Research and Development, 2019, 56(10): 2071−2096 (in Chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071−2096)

[9] Yang Pengbo, Sang Jitao, Zhang Biao, et al. Survey of the interpretability of deep models for image classification[J]. Journal of Software, 2023, 34(1): 230−254 (in Chinese)
(杨朋波, 桑基韬, 张彪, 等. 面向图像分类的深度模型可解释性研究综述[J]. 软件学报, 2023, 34(1): 230−254)

[10] Chatonsky G. Deep dream (The Network's Dream)[J]. SubStance, 2016, 45(2): 61−77

[11] Abbasi A R, Yu Bin. Interpreting convolutional neural networks through compression[J]. arXiv preprint, arXiv: 1711.02329, 2017

[12] Li Yuchao, Lin Shaohui, Zhang Baochang, et al. Exploiting kernel sparsity and entropy for interpretable CNN compression[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2800−2809

[13] Mike W, Michael C H, Sonali P, et al. Beyond sparsity: Tree regularization of deep models for interpretability[C] //Proc of the AAAI Conf on Artificial Intelligence Menlo Park, CA: AAAI, 2018: 1670−1678

[14] Himabindu L, Ece K, Rich C, et al. Interpretable & explorable approximations of black box models[J]. arXiv preprint, arXiv: 1707.01154, 2017

[15] Tan S, Caruana R, Hooker G, et al. Learning global additive explanations for neural nets using model distillation[J]. arXiv preprint, arXiv: 1801.08640, 2018

[16] Krishnan R, Sivakumar G, Bhattacharya P. Extracting decision trees from trained neural networks[J]. Pattern Recognition, 2019, 1(32): 12

[17] Yang Chengliang, Anand R, Sanjay R. Global model interpretation via recursive partitioning[C] //Proc of the IEEE 20th Int Conf on High Performance Computing and Communications; IEEE 16th Int Conf on Smart City; IEEE 4th Int Confon Data Science and Systems. Piscataway, NJ: IEEE, 2018: 1563−1570

[18] Fan Fenglei, Wang Ge. Fuzzy logic interpretation of quadratic networks[J]. Neurocomputing, 2020, 374: 10−21

[19] Carvalho D V, Pereira E M, Cardoso J S. Machine learning interpretability: A survey on methods and metrics[J]. Electronics, 2019, 8(8): 832

[20] Samek W, Montavon G, Lapuschkin S, et al. Toward interpretable machine learning: Transparent deep neural networks and beyond[J]. arXiv preprint, arxiv: 2003.07631, 2020

[21] Liu Junhong , Lin Yijie, Jiang Liang, et al. Improve interpretability of neural networks via sparse contrastive coding[C] //Proc of the 2022 conf on Empirical Methods in Natural Language Processing (EMNLP). Abu Dhabi, United Avab Emirates, 2022: 460−470

[22] Ribeiro M T, Singh S, Guestrin C. "why should I trust you?": Explaining the predictions of any classifier[C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135−1144

[23] Vitali P, Abir D, Kate S. RISE: Randomized input sampling for explanation of black-box models[C] //Proc of the British Machine Vision Conf (BMVC). Newcastle, UK: BMVA, 2018: 151

[24] Harini S, Nathan H, Alistair J, et al. Clinical intervention prediction and understanding using deep networks[J]. arXiv preprint, arXiv: 1705.08498, 2017

[25] Christoph M, Giuseppe C, Bernd B. Interpretable machine learning—A brief history, state-of-the-art and challenges[C] //Proc of the Workshops of the European Conf on Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Berlin: Springer, 2020: 417−431

[26] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net[J]. arXiv preprint, arXiv: 1412.6806, 2014

[27] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-Linear classifier decisions by layer-wise relevance propagation[J]. PLOS ONE, 2015, 10(7): 130140

[28] Avanti S, Peyton G, Anshul K. Learning important features through propagating activation differences[C] //Proc of the Int Conf on Machine Learning. New York: International Machine Learning Society (IMLS), 2017: 3145–3153

[29] Zhang Jianming, Sarah A B, Zhe Lin, et al. Top-down neural attention by excitation backprop[J]. International Journal of Computer Vision, 2018, 126(10): 1084−1102

[30] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2921–2929

[31] Selvaraju R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C] //Proc of the IEEE Conf on Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 618–626

[32] Wang Haofan, Du Mengnan, Yang Fan, et al. Score-cam: Improved visual explanations via score-weighted class activation mapping[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 111–119

[33] Jeong R L, Sewon K, Inyong P, et al. Relevance-cam: Your model already knows where to look[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 14944−14953

[34] Fan Fenglei, Xiong Jinjun, Li Mengzhou, et al. On interpretability of artificial neural networks: A survey[J]. IEEE Transactions on Radiation and Plasma Medical Sciences, 2021, 5(6): 741−760

[35] Chen Chaofan, Oscar L, Daniel T, et al. This looks like that: Deep learning for interpretable image recognition[C] //Proc of the 33rd Conf on Advances in Neural Information Processing Systems. Cambridge, MA: Neural Information Processing Systems Foundation, 2019: 8930–8941

[36] Oscar L, Liu Hao, Chen Chaofan, et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions[C] //Proc of the AAAI Conf on Artifcial Intelligence, Menlo Park, CA: AAAI, 2018: 3530−3537

[37] Peng Xi, Li Yunfan, Tsang I W, et al. XAI Beyond classification: Interpretable neural clustering[J]. Journal of Machine Learning Research 2022, 23: 6: 1−6: 28

[38] Zhang Quanshi, Wu Yingnian, Zhu Songchun. Interpretable convolutional neural networks[C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8827−8836

[39] Lage I, Ross A, Gershman S J, et al. Human-in-the-loop interpretability prior[C] //Proc of the Neural Information Processing Systems. Cambridge, MA: Neural information processing systems. foundation, 2018: 10159−10168

[40] Du Jian, Zhang Shanghang, Wu Guanhang, et al. Topology adaptive graph convolutional networks[J]. arXiv preprint, arXiv: 1710.10370, 2017

[41] Wen Zaiwen, Yin Wotao. A feasible method for optimization with orthogonality constraints[J]. Mathematical Programming, 2012, 142: 397−434

[42] Christopher M, Martin R, Matthias F, et al. Weisfeiler and Leman go neural: Higher-order graph neural networks[C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 4602−4609

Zhao Xiaoyang, born in 1999. Master candidate. Her main research interests include deep learning and interpretability of deep neural networks.

赵小阳, 1999 年生. 硕士研究生. 主要研究方向为深度学习和深度神经网络的可解释性.

Li Zhongnian, born in 1990. PhD, lecturer. His main research interests include machine learning, data mining, and medical image processing.

李仲年, 1990 年生. 博士, 讲师. 主要研究方向为机器学习、数据挖掘和医学图像处理.

Wang Wenyu, born in 1999. Master candidate. Her main research interests include deep learning and interpretability of deep neural networks.

王文玉, 1999 年生. 硕士研究生. 主要研究方向为深度学习和深度神经网络的可解释性.

Xu Xinzheng, born in 1980. PhD, professor, PhD supervisor. Member of CCF. His main research interests include machine learning, data mining, and medical image processing.

许新征, 1980 年生. 博士, 教授, 博士生导师. CCF 会员. 主要研究方向为机器学习、数据挖掘和医学图像处理.