基于可解释基拆解和知识图谱的深度神经网络可视化

阮利[1),2)] 温莎莎[2)] 牛易明[2)] 李绍宁[2)] 薛云志[3)] 阮涛[4)] 肖利民[1),2)]

[1)] (软件开发环境国家重点实验室北京 100191)

[2)] (北京航空航天大学计算机学院北京 100191)

[3)] (中国科学院软件研究所北京 100190)

[4)] (中国专利信息中心北京 100088)

摘 要 近年来，以卷积神经网络（CNN）等为代表的深度学习模型，以其深度分层学习，无标签化学习等优势，已在图像识别为代表的各个领域得到日益广泛的应用。然而，深度神经网络模型由于其内在的黑盒原理，对其内部工作机制的解释仍然面临巨大挑战，其可解释性问题已成为了研究界和工业界的前沿性热点研究课题。针对现有研究存在的缺乏基于图谱的可解释性方法的问题，以及可解释基模型的图谱构建优势，本文提出了一种基于可解释基拆解和知识图谱的深度神经网络可视化方法。首先采用一种面向可解释基模型特征拆解结构的知识图谱构建方法，构建了场景和解释特征之间的解释关系和并列关系等图谱信息；利用场景-特征的解释关系网络，提出了一种基于 Jaccard 系数的场景间相似度聚类方法；针对现有可解释基模型对相似的场景，其解释特征重合率可能很高的问题，提出了一种基于场景的判别性特征提取方法，在特征拆解结果中能对每一类样本分别提取出了能够区别此类和其他类并且拥有同等重要性的拆解特征（即判别性特征）；针对现有可解释基的深度网络可视化测试缺乏保真度测试的问题，提出了一种适于深度神经网络的保真度测试方法。保真度测试和人类置信度测试，均表明本文所提方法可取得优异效果。

关键词 深度神经网络；可视化；可解释基拆解模型；知识图谱；解释深度学习模型

中图法分类号 TP391

Deep Neural Network Visualization Based on Interpretable Basis Decomposition and Knowledge Graph

Li Ruan[1),2)] Shasha Wen[2)] Yiming Niu[2)] Shaoning Li[2)] Yunzhi Xue[3)] Tao Ruan[4)] Limin Xiao[1),2)]

[1)] (State Key Laboratory of Software Development Environment，Beijing 100191)

[2)] (School of Computer Science and Engineering, Beihang University, Beijing 100191)

[3)](Institute of Software, Chinese Academy of Science, Beijing 100191)

[4)](China Patent Information Center, Beijing 100088)

Abstract Recently, owing to the advantages of deep-layered learning and unlabeled learning, etc., deep learning models represented by convolutional neural network, deep neural network, recurrent neural network, have gained increasing applications in various fields, such as image recognition, video, and natural language processing. To achieve the high transparency and security assurance of deep learning models, the interpretability research of deep neural networks is of great theoretical significance and industrial application value and recently gains increasingly attentions. However, because of the intrinsic black-box characteristics of the deep learning models, the interpretation of its internal structure and the running mechanism is still of great challenges, including the rigorous theo-

本课题得到国家重点研究计划（NO. 2017YFB0202004）、软件开发环境国家重点实验室课题（ No. SKLSDE-2020ZX-15）和国家自然科学基金青年项目（No.11701545,No.61772053）资助。阮利，博士，硕士生导师，研究方向为AI安全、时序分析、知识图谱和分布式系统。E-mail:ruanli@buaa.edu.cn。温莎莎，学士，研究方向为知识图谱。牛易明，本科生，研究方向为知识图谱。李绍宁，本科生，主要研究方向为时序分析，网络安全。薛云志，博士，研究员，主要研究方向为可信赖人工智能、人工智能测试与评估、知识图谱。阮涛，硕士，助理研究员，研究方向为国内外专利翻译等。肖利民，博士，教授，研究方向为高性能计算、分布式系统。

retical results originated from the manual observations of large-scale training and testing set, and scarce appropriate explanation of the learning results based on the human understanding. Moreover, most of the existing researches analyzing the decision-making process of deep learning models only from a local perspective and lacks a graphical representation based on the overall understanding. On the other hand, the interpretable basis decomposition (IBD) model has the advantages that its interpretation result is not only a strict corresponding relation from scene to feature, but also is a kind of semi-structured data which can facilitate IBD based knowledge map construction from it. Aiming at the problem that existing deep neural network visualization researches lacks the interpretability based on the knowledge map and the well-suited knowledge map representability of IBD, we propose a deep neural network visualization approach based on interpretable basis decomposition and knowledge map, which fully takes the advantage of map construction ability of interpretable basis decomposition. Firstly, we propose a knowledge map construction method based on the feature decomposition structure of IBD, which constructs the map information, such as the interpretation relationship and juxtaposition relationship, between the scene and the interpretable feature. Then, a similarity clustering algorithm between scenes using Jaccard coefficient based on the interpretation relation network of scenes and features is proposed. Based on a scene discriminant feature extraction method, the decomposed features that can distinguish this class from other classes are extracted from each type of sample, namely discriminant features. Meanwhile, we quantify the accuracy of discriminant feature extraction by means of manual evaluation by exploring the difference between different models' understanding of the recognition target and that of human beings. Furthermore, a fidelity test method for deep network has been proposed to solve the problem that existing research lacks fidelity test. We combine the multi-feature thermal spectrograms into a comprehensive characteristic thermal spectrogram, and then use the Hadmag product to refuse the comprehensive characteristic thermal spectrogram with the original image to obtain the characteristic fusion spectrogram. The luminance labeled depth neural network classification model of feature fusion map was used to identify the target location pixel area, and the target location ability of thermal spectrum map was measured by comparing the deviation of input original map and feature fusion map to the model classification ability, so as to obtain the fidelity of the interpretable basis decomposition model. Both the fidelity test and the human confidence test show that the proposed method can achieve excellent results.

Key words Deep Neural Networks; Visualization; Interpretable Basis Decomposition; Knowledge Map; Interpreting Deep Learning Models

1 引言

近年来，以深度神经网络（Deep Neural Network, DNN）、卷积神经网络（Convolutional Neural Network, CNN）[1]、循环神经网络（Recurrent neural network,RNN）**Error! Reference source not found.**等为代表的深度学习模型以其深度分层学习、无标签化学习模型等优势，已在计算机视觉、语音识别、自然语言处理为代表的各个领域都得到日益广泛的应用。

近年来，随着深度学习应用领域的不断拓展，作为制约深度学习更深入和长远应用的瓶颈，可解释性问题受到各领域研究者的日益广泛的重视[3][4]。深度学习的模型学习、训练的推演过程以数字运算为主导且具有黑盒性质，缺乏可解释性且难以通过人类社会的概念进行理解。更具体的挑战包括：(1)在深度学习观测值和结果层面，由于对观测值无法进行严密的逻辑推理，进而观测结果缺乏强有力的理论支撑或基于现实的合理解释。(2)在深度学习网络层面，深度网络缺乏因果逻辑推理，因此深度神经网络存在一定安全隐患，难以对某些神经网络的木马攻击进行有效防范，如特洛伊木马[5]。可见，如何提高深度神经网络模型的透明性已成为了当前一个前沿和热点的研究课题。

神经网络的可解释性研究经历了从早期利用了模型对输入的敏感度差异性特征的粗粒度探索，到近期探索单个/组合神经元在一次识别任务中的功能体现的过程。然而现有研究的思路仍然仅仅是从局部的角度分析神经网络的决策依据/决策过程，缺乏对神经网络学习和决策的整体理解的图形化的抽象表示，尤其缺乏一套完整的神经网络学习结果的知识图谱构建方法和实例。因此，针对现有研

究存在的上述问题，开展深度神经网络的可解释性研究将对深度网络模型透明性、安全性保障都有重大的理论研究意义和工业应用价值。另一方面，由于可解释基拆解模型对深度神经网络的解释结果为场景到特征的严格对应关系，是一种半结构化数据，从结构上就具有便于进行知识图谱构建的优势。因此，针对现有深度神经网络研究缺乏基于可解释基拆解和图谱融合的可解释性方法的问题，以及可解释基模型的图谱构建优势，本文提出了一种基于可解释基拆解模型和知识图谱的深度神经网络可视化方法。

2 相关研究

从可解释性的角度给机器学习模型分类，主要有两大类：自解释模型和依赖外部可解释性技术的模型。

深度神经网络属于后者——依赖外部可解释性技术的模型，针对深度神经网络的解释算法也可从多种角度进行设计。下文将按规则提取、显著性映射、深度网络表示这三大解释角度对上述两种模型分析最新的研究进展及其不足。

2.1 基于规则提取的可解释性相关研究

规则提取是研究人员最早提出的对黑盒模型的解释方法，其把已训好的模型当作黑盒，借助自解释模型的优势，使用自解释模型模拟黑盒模型的决策过程，抽象出一系列决策依据，使得自解释模型尽可能接近黑盒模型的决策能力，这样的自解释模型也被称为代理模型。代理模型主要有两大类：线性代理模型和决策树模型。

2016 年，Ribeiro 提出的线性代理模型 LIME[6] 通过探测输入数据扰动构建一个局部线性模型，用于判断输入数据中某些部分对模型输出结果的影响度，且可根据线性模型非零维数量来量化表示 LIME 解释性的复杂度。2016 年，Jan 提出 DeepRED 模型[7]构建的决策树几乎达到了深度神经网络模型的完整性，但执行时间和内存开销大是其一大弊端。2019 年 UCLA 的张拳石教授团队提出一种新的决策树解释模型[4]，该模型挖掘了 CNN 的所有潜在决策模式，决策模式提供从粗到细粒度的决策依据，用以解释 CNN 在不同粒度上的预测依据。然而，基于规则提取的方法存在依赖于已训练模型、泛化性和可扩展性受限等问题。

2.2 基于显著性映射的可解释性相关研究

2011 年，纽约大学的 Zeiler[8]等人提出了一种通过卷积稀疏编码和最大池交替层学习图像分解的分层模型 Adaptive Deconvnet。2013 年，Zeiler[9]通过对 CNN 逐层还原结果表明，从 Deconvnet 模型的深浅层可提取图像不同粗细粒度的信息，以及模型对输入图像的平移、缩放不敏感，但对图像的旋转敏感。但以上工作忽略了神经网络内部的除了梯度以外的其他重要信息。其他显著性映射方法还有：LRP、DeepLIFT[10]、CAM[11]、Grad-CAM[12]、Grad-CAM++[13]、Integrated gradients[14]等。这些工作核心思想是利用神经元激活值找到输入样本中对输出结果影响最大的区域，以及高网络敏感度。ETH 的 Marco Ancona[15]等人对比了以上显著性映射方法的解释能力。

2015 年，法国国家信息与自动化研究所的 Oquab[16] 提出使用全局最大池化（global max-pooling）方式，对目标分类模型的识别点进行定位。受[16]启发，2016 年，MIT 的 Bolei Zhou[11] 等人提出一种类映射激活（CAM）方法，用于具有全局平均池化（GAP）的 CNN 模型，并证实 CNN 提取的特征含有位置信息。2018 年，MIT 的 Bolei Zhou[3] 在之前的研究[11]基础上，提出一种 Interpretable Basis Decomposition（IBD）方法，核心思想是对 CNN 得到的 activation map（AM）进行解码得到对预测结果的合理解释。2018 年，Selvaraju[17]利用神经元能够在分类任务中从训练集中学习到的物体特征的特性，可以通过神经网络可视化的技术获取到模型学习到的概念，用以构建相关领域的知识网络。

2.3 基于深度网络表示的可解释性相关研究

深度网络表示的出发点是从网络结构本身的功能来解释网络的决策依据。

深度网络表示的工作可按网络结构层次分为：层级解释、神经元级解释、向量级解释。2014 年，Razavine[18]发现，使用 ImageNet 数据集训练的分类网络的内部层的输出产生了一个特征向量，可通过复用该特征向量来解决对不同种类的鸟类进行细粒度分类、属性检测和对象定位等图像处理问题。2014 年 Nguyen[19]等人对 AlexNet 模型进行实验，使用梯度上升法最大化 softmax 输出，最终 DNN 分类模型对图像的识别结果可信度达 99.99%，在 MINIST 数据集上错误率达 0.94%。

康奈尔大学的 Yosinski[20]等人在 2015 年提出并实现了两种不同的神经网络可视化工具，最终得到辨识度更高的图像。

MIT 的 Bolei Zhou 等人在 2017 年提出 Network Dissection 模型[21]，模型通过评估隐层神经元和一系列语义概念之间的契合度来解释神经网络，与实际意义关联度高的神经元被赋予具象化解释，如物体标签、场景的具体某个部分、文字、材料和颜色等，并以一系列神经元的随机线性组合为单位，赋予网络实际意义。同时提出了“分割表示”的概念，借鉴独热编码的思想，通过将神经网络黑盒学习的特征分割成若干个人类可辨识的视觉概念特征。基于 Bolei Zhou[21]的分割思想，2019 年，MIT 的 David Bau**Error! Reference source not found.**等人将 CNN 的分割延展到 GAN 模型上，这篇论文介绍了一种对 GAN 模型的可视化框架，通过人为定义一些检测功能神经元，并将这一改动介入到网络中，探索 GAN 网络的人为介入对模型的影响，进而推测并解释 GAN 模型。Cao 等[23]采用观察神经网络的反馈来分析 CNN 的视觉定位与分割。近几年，有研究人员探究单个神经元的线性组合在表示空间中的其他方向形成的表示，如 2018 年谷歌大脑提出的概念激活向量解释方法（Concept Activation Vectors, CAVs）**Error! Reference source not found.**。

可解释基拆解模型（IBD，Interpretable Basis Decomposition，以下简称 IBD 模型）[3]是 MIT 的 Bolei Zhou 等人近年来新提出的对 CNN 的另一种可视化方法。算法核心思想是拆解 CNN 最后一层激活特征向量，将多分类任务中对每个识别目标的激活特征向量拆解，最终分解成若干个不同相对更细粒度概念特征向量的表示。然而，以 IBD 模型的直接结果（如百分比方式）对网络结构的解释缺乏场景类型维度的抽象，只能提取出单个输入样本的特征解释结果，而无法对一类场景或整个数据集样本总体进行解释。同时[3]对可解释基拆解模型的测试仅覆盖人类置信度的度量，缺乏保真度做量化测试，模型存在测试维度不全面的问题。本文的工作针对 Bolei Zhou[3]的工作存在的结果采用百分比，缺乏场景理解和以及 IBD 测试存在的问题，提出了新方法。

综上可知，虽然知识图谱具有更直观的解释能力，是近年来进行可视化直观解释的前沿技术，然而深度学习神经网络可解释性的现有研究中，仍然缺乏一套完整的神经网络学习结果的知识图谱构建方法和实例。另一方面，虽然现有的研究已经有基于可解释基拆解模型的深度可视化方法，但是现有方法并没有有效利用到可解释基拆解模型对深度神经网络的解释结果为场景到特征的严格对应关系，是一种半结构化数据，其结构上就具有便于进行知识图谱构建的优势。

3 基于可解释基拆解和知识图谱的深度网络可视化建模及问题分析

基于可解释基拆解的深度神经网络可视化方法，主要针对场景是：在 CNN 等分类模型中，有不同的网络结构实现，如 Resnet18、Resnet50、VGG、AlexNet 等深度网络结构在同一数据集上训练的识别效果不同。因为神经网络的黑盒特性，所以模型应用人员无法直接分辨是什么原因导致这些模型在同样的数据上有不同的识别效果。IBD 方法的目的就是为了探究对同一识别目标的识别任务中，这些深度网络学习了哪些特征，以及哪些特征更能作为识别目标的特有属性，在多分类任务中帮助区分不同识别目标。IBD 深度网络可视化方法主要是通过作用于 CNN 最后一层的激活特征向量，最终得到每类识别目标对应激活特征向量的具象化概念标签拆解，并给出百分比的形式衡量具像化特征在 CNN 识别模型中的重要性。向量拆解表示和概念特征热谱图的构建算法是本文 IBD 模型中的核心，下面首先引入 MIT[3]中对这两者的理论介绍，为本文的算法提出打下理论基础。

3.1 理论基础

3.1.1 深度网络向量拆解表示

假设$f(x) \in R^K$为深度网络对输入x的K维输出结果。f_k(x)表示输入x对应分类结果为k的概率，由此可得输入标签为c的x样本被误分到k标签的概率$f_k(x)$。用$h(g(x))$作为f(x)的中间表示方式，其中$h(a)$为网络最顶层，$a = g(x) \in R^D$为表示域中的一个点。在本算法中，将 CNN 倒数第二层输出抽象为$a = g(x)$，$h(a)$为简单线性组合，h_k可表示为w_k和a的线性组合，如公式（1）所示。

$$h(a) = W^{(h)}a + b^{(h)}, \#(1)$$
$$h_k(a) = w_k^T a + b_k$$

假设有一系列向量$q_{c_i} \in R^D$，每个概念特征c_i，总存在一个向量q_{c_i}与之对应。其中c_i标签比k标签粒度更细，用于辅助解释分类结果。这样w_k可向量

分解为公式（2）所示，这样的一组q_{c_i}就是概念特征正交基。

$$w_k \approx s_{c_1} q_{c_1} + \cdots + s_{c_n} q_{c_n} \#(2)$$

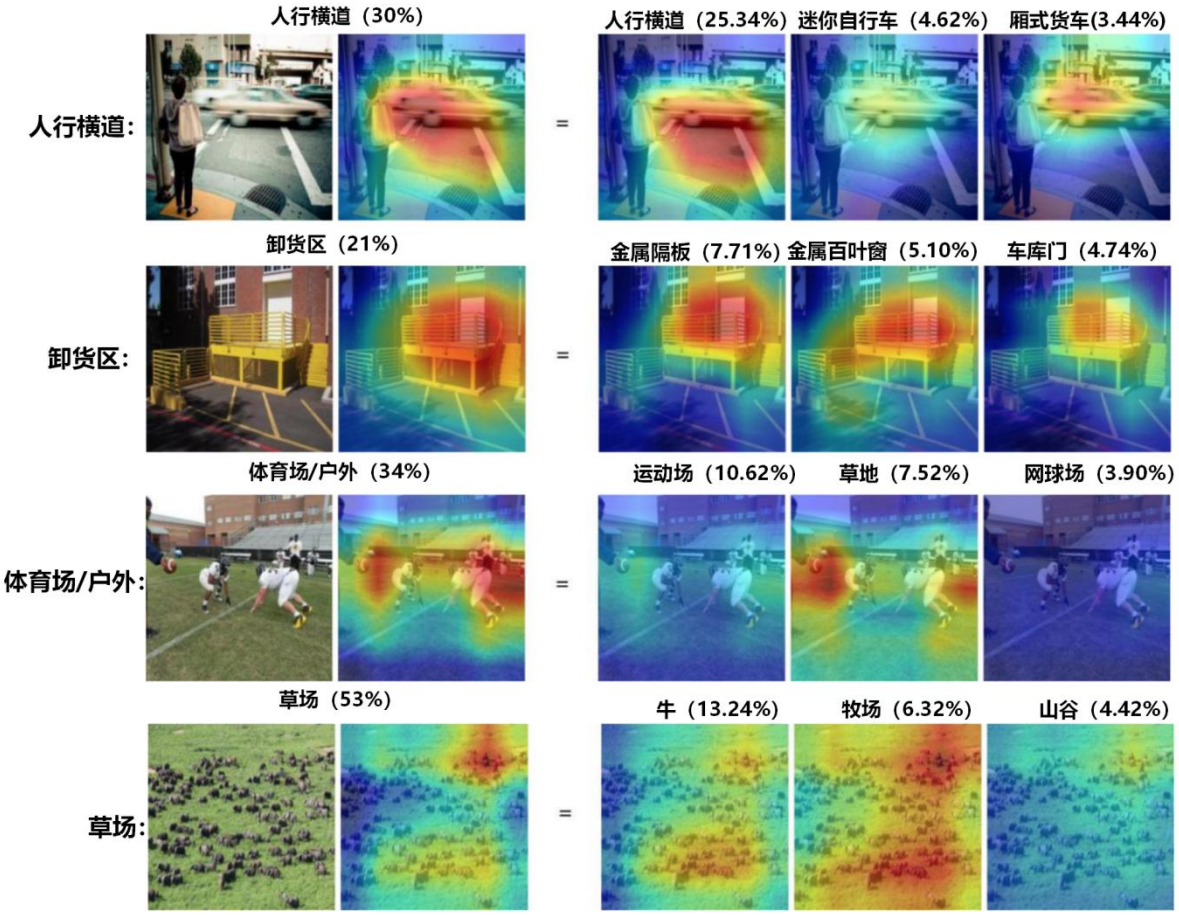

图 1 IBD 模型在 Resnet18 网络结构和 Places365 数据集上的类向量拆解结果可视化。图示为其中 4 种识别场景，如第一个例子中，识别对象为人行道，拆解结果中重要性百分比排前三的特征分别是：人行道、迷你自行车和厢式货车，重要性分别占 25.34%、4.62%和 3.44%。在热谱图中，网络在红色部分（暖色）活性较高，蓝色部分（冷色）活性较低，即网络更关注热谱图中活性高的区域。

将$\mathrm{q}_{\mathrm{c_i}}$用C矩阵表示，则使得该式子两边最接近的问题可类比为最小二乘问题，如公式（3）所示。即找到$s_{\mathrm{c_i}}$使得$|r|$最小。

$$\begin{aligned} w_k &= s_{c_1} q_{c_1} + \cdots + s_{c_n} q_{c_n} + r \#(3) \\ &= Cs + r \end{aligned}$$

由此得到的$s = C^{+} w_k$为最优解，其中C^{+}是C的伪逆解。

3.1.2 概念特征热谱图

在图像识别任务中，由于图像包含的信息非常丰富，尽管深度网络用到的数据集(如 Broden 数据集等）已经包含丰富且密集的标签，但也无法囊括 CNN 在图像识别任务中提取到的特征。因此本文考虑在候选基向量组C_k中加入一项残差向量$r = w_k - C_k s$，记为C_k^*。这样C_k^*包括了整个 CNN 学习到的特征表示。对于 CNN 最后一层分类结果为k的分数可表示为如公式（4）所示。

$$h_k(a) = w_k^T a + b_k$$

$$= (C_k^* s)^T a + b_k \#(4)$$
$$= s_1 q_{c_1}^T a + \cdots + s_n q_{c_n}^T a + r^T a + b_k$$

其中$s_1 q_{c_i}^T a$为拆解向量对应概念标签c_i对识别结果为k的贡献度，$r^T a$为残差r对识别结果为k的贡献度（可理解为非 Broden 数据集包含的标签）。

因此对每个标签c_i可作用于池化层，得到输入图像x关于特征c_i的热谱图如公式（5）所示。

$$s_i q_{c_i}^T a = s_i q_{c_i}^T pool(A) \#(5)$$
$$= pool\left(s_i q_{c_i}^T A\right)$$

其中$q_{c_i}^T A$为特征c_i的热谱图。

3.2 IBD模型解释结果可视化效果及问题分析

以 Resnet18 网络结构为例，我们运用 IBD 模型在 Places365 数据集上对其进行解释，解释结果可视化效果如图 1 所示，图中展示了样本对应的类向量特征热谱图，以及该类向量拆解得到的一组概念特征正交基的贡献度最大的 4 组热谱图。图 1 为其中 4 种识别场景，如第一个例子中，识别对象为人行道，拆解结果中重要性百分比排前三的特征分别是：人行道、迷你自行车和厢式货车，重要性分别占 25.34%、4.62%和 3.44%。在热谱图中，网络在红色部分（暖色）活性较高，蓝色部分（冷色）活性较低，即网络更关注热谱图中活性高的区域。

以上解释结果可视化图是 IBD 模型的直接解释结果，主要以两种形式呈现：特征热谱图和特征贡献度百分比。从中我们只能直观地感受场景的主要特征组成，以及从热谱图中可以直接看出这些特征在原图中的像素分布区域。可见，由于以上两点均是以输入样本为解释单位，缺乏对数据集中场景类型维度的解释。模型对结果的解释为从拆解特征中提取出权重最高的几个特征作为类的解释，对解释结果的分析角度较为单一，这些权重是仅针对一个样本而言，并不能体现该类所有样本的普遍性结果，权重仅能体现某一特征对分类目标的重要性，而不能体现该特征是否有助于区别其他类特征。

即以 IBD 模型的直接结果对网络结构的解释缺乏场景类型维度的抽象，只能提取出单个输入样本的特征解释结果，而无法对一类场景或整个数据集样本总体进行解释。我们的研究目标是希望解释模型能够达到以场景为单位的解释效果，对一类场景的特征做更进一步的可视化工作，进而构建出对整个数据集样本的抽象解释结果。

4 基于可解释基拆解和知识图谱的深度神经网络可视化

4.1 基本思路和总体设计

基于可解释基拆解和知识图谱的深度神经网络可视化方法核心思想是：基于第 3 节理论，首先采用 IBD 模型通过作用于 CNN 最后一层的激活特征向量，最终得到每类识别目标对应激活特征向量的具象化概念标签拆解；然后通过知识图谱更加图谱化和场景化地衡量具像化特征在 CNN 识别模型中的重要性，将可以直观地看出神经网络对不同标签事物的决策依据，知识图谱内容在训练数据集包含的范围内，既能对不同的数据集可产生不同的知识图谱，也可在训练时融合多个数据集的知识或分别在不同的数据集上训练模型并把解释结果融合成一个覆盖范围更广的图谱。

基于上述设计思路，本文方法的模型（图 2）

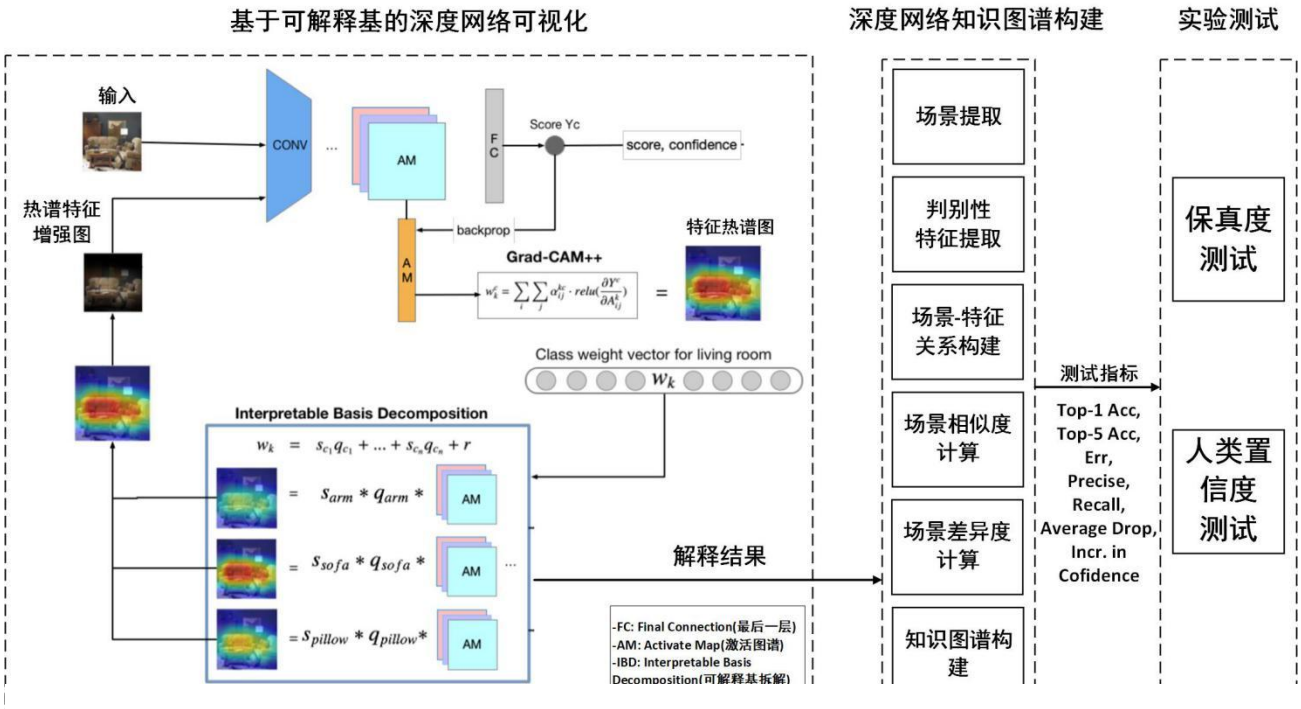

图 2 基于 IBD 和知识图谱的深度神经网络可视化算法总体设计

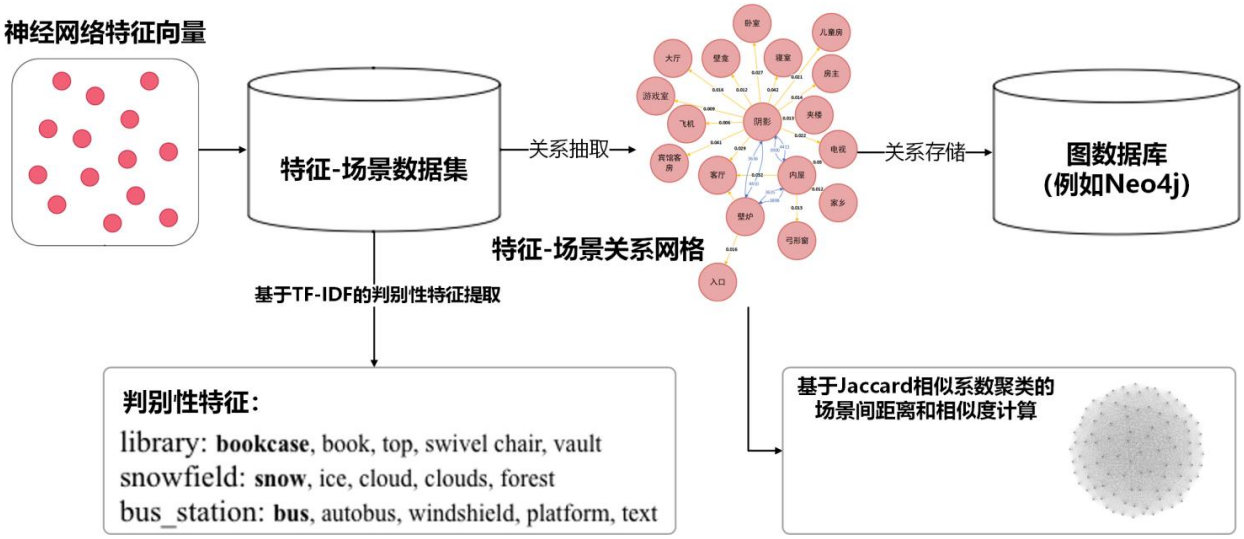

图 3 IBD 模型对 resnet18 网络结构的解释结果，在数据集 places365_val 上的知识图谱构建总体设计图

主要包括 3 个核心：（1）基于 IBD 对深度神经网络进行可视化；（2）可解释基拆解模型特征拆解结果的知识图谱构建。（3）基于保真度和人类置信度的测试。总体设计如图 2 所示。如图 2 所示，其中“（1）基于 IBD 对深度神经网络进行可视化方法”的网络结构构成方法是包括神经网络的输入数据集（图 2 左上角），改进的 CNN 网络结构（图 2 中间部分）。其中改进 CNN 网络结构主要指将 CNN 最后一层全连接层替换为 GAP 层。其算法流程是，首选通过可解释基拆解算法将权向量拆解为若干个特征向量的表示。然后，对每个特征向量，使用 Grad-CAM 算法（用每一类的分数对倒数第二层激活图进行一次反向传播求导），可以得到特征热谱图。

其中“（2）可解释基拆解模型特征拆解结果的知识图谱构建”方法主要对于现有对 IBD 模型的结果解释缺乏对神经网络学习和决策的整体抽象问题，本文通过场景提取、判别性特征提取、场景特征关系构建、场景相似度和差异度计算步骤，构建了一个完整的神经网络学习结果的知识图谱的方式，实现对 IBD 模型解释结果的知识图谱化解释。

作为示例，图 3 以 IBD 模型对 resnet18 网络结构的解释结果为例，展示了本文方法在数据集 places365_val 上的知识图谱构建步骤和效果。图 3 中输入部分为 IBD 模型解释结果构成的特征-场景数据集，特征和场景被视为同类节点；中间部分为通过关系抽取构建的特征-场景关系网络，该关系将被存储到 Neo4j 图数据库中；针对特征-场景之间的解释关系和加权解释关系分别进行本文的场景间相似度计算和基于 TF-IDF (Term Frequency-Inverse Document Frequency) 指标提取每类样本的判别性特征提取。更详细的关键技术将在后续详细分析。

4.2 算法流程

特征向量确定和图谱输入是基于 IBD 和图谱模型进行深度神经网络可视化的输入的关键。本文算法输入模型的特征向量来自训练数据集，使用 IBD 和知识图谱模型对其验证数据集在多种深度神经网络结构上的多分类任务进行解释。现有的深度学习网络的训练数据集常常是融合了多个不同标注的数据集，例如 ADE、OpenSerfaces、Pasal-Context、Pasca-Part 和 Describale Textures DataSet，每个样本都有若干个 pixel-wise 的标签（例如物体、场景、物体的部分、纹理、材料和颜色），除了纹理和场景标签外，大部分样例都细分到像素级别，每张图片都对应若干个标签图（label map）。用于构建知识图谱的数据包含两大类概念：类别和特征，数据以文本标签的形式存储。IBD 解释模型的拆解结果为半结构化数据，本文选取每一类样本拆解向量平均贡献度 Top5 的特征构建知识图谱，并将特征和类别视为概念，而非从属关系，从而使数据结构化。下面进一步结合图 2，介绍本文的算法流程（以“客厅”场景为例）。

基于可解释基拆解和知识图谱的深度神经网络可视化算法的核心步骤设计如下所示（以“客厅”为例，目标是：输入“客厅”场景，本文方法能够输出与“客厅”场景相关的场景-特征关系网络。

1. 场景识别：将场景图片（如“客厅”）输入场景识别深度网络模型（如 CNN），得到判别结果为

“客厅”的权向量w_k、打分$score$、置信度$confidence$；

2. 可解释基拆解：可解释基拆解算法将 CNN 判别结果为“客厅”的权向量拆解为若干个特征向量的表示，得到的拆解结果严格依赖于可解释基拆解模型的训练数据。如图 2 所示“客厅”被拆解为：“扶手”、“沙发”、“靠枕”这三个特征向量；

3. 输出特征热谱图：为了达到更直观的可视化效果，以及方便后续对拆解结果准确性校验，对上一步得到的每个特征向量在原图中进行可视化，对 CNN 最后一层激活图使用 Grad-CAM 算法，经过一次反向传播求导得到对应的特征热谱（图 2）；

4. 进行特征-场景解释：基于哈德玛积输出场景的热谱（mask）图，我们使用重要程度最高的三个特征作为“客厅”场景的解释：“扶手”、“沙发”、“靠枕”，合并三个特征热谱图得到“客厅”场景的热谱图，然后进行特征-场景解释；

5. 场景热谱增强图生成：基于哈德玛积在原图中对这三个特征进行增强处理，生成“客厅”场景的 mask 图；

6. 计算场景识别效果评分：这一步作为可解释基拆解模型的算法保真度测试，将上一步得到的 mask 图重新输入到第一步的场景识别 CNN 模型，得到打分$score_{mask}$、置信度$confidence_{mask}$。通过生成热谱图对应的 mask 图，对比实验测试 mask 集和原图集在分类神经网络中的准确率等指标，可量化测试热谱图在目标识别任务下的聚焦能力，并分析网络结构在物体识别任务上的可优化点。

7. 计算可解释基拆解模型保真度：对原图的打分$score$、置信度$confidence$和 mask 图的打分$score_{mask}$、置信度$confidence_{mask}$数据进行分析，得到可解释基拆解模型保真度的量化结果；

8. 构建场景知识图谱：利用第二步得到的半结构化数据（包含场景、拆解特征以及场景-特征之间的关系）构建知识图谱，并把知识图谱存储到图数据库。以“客厅”场景为例，第二步得到三个特征向量“扶手”、“沙发”、“靠枕”，利用三个向量对“客厅”场景的解释关系，构建知识图谱。这一步可扩展性和灵活性很强，我们根据 IBD 模型的训练集，可以解释在训练集范围内或训练集足以作为特征向量表示的一类场景，最终构建出特定类型场景的知识图谱；

9. 场景间相似度计算：这一步和下一步将作为对解释结果的进一步延伸。上一步我们构建了场景-特征关联关系，包括场景和解释特征之间的解释关系和并列关系等信息，在此基础上，我们采用基于 Jaccard 相似系数的场景间相似度算法计算场景间相似度。在“客厅”例子中，关系包括：“扶手”、“沙发”、“靠枕”对“客厅”的解释关系，以及“扶手”、“沙发”、“靠枕”两两间的并列关系。在第 4 章中，我们将会以“客厅”为例，探讨如何利用 Jaccard 相似系数计算两个场景间的相似度。

10. 提取判别性特征：这一步作为解释结果的进一步延伸，对 IBD 模型特征拆解结果使用 TF-IDF 指标提取每一类的判别性特征。以“客厅”场景为例，第二步得到的三个特征向量“扶手”、“沙发”、“靠枕”，使用 TF-IDF 指标最高的特征“沙发”作为“客厅”场景的判别性特征；

11. 人类置信度测试：对上一步判别性特征进行人类置信度测试，基于判别性特征的 TF-IDF 值计算 MSE 指标可量化表示人类置信度。

4.3 场景-特征关系建模

本节首先介绍深度学习模型结果的场景和拆解特征进行关系抽取，构建场景-特征的知识图谱，并存储到 Neo4j 图数据库的方法。然后使用 TF-IDF 指标对模型解释结果提取每一类场景的判别性特征，同时提出一种场景间距离和相似度计算方法，并用距离公式对深度学习训练数据集进行聚类分析和可视化，算法由于针对特征拆解结果设计，可以直接反映出深度神经网络经过迭代训练后对数据集不同场景的理解，且是基于训练集和网络结构两个维度上的相似度计算方法，这种相似度计算以图像的形式理解人类社会的概念，优势是可以跨越语言的差异，以绝对的事物概念理解不同场景，另外，这种相似度表示方式与人类理解事物的方式接近，即这种相似度计算公式容易加入人类的其他外部知识，可扩展性强，

4.3.1 场景-特征关系抽取算法

场景-特征的关系抽取核心是从数据集中提取出场景和特征之间的关联关系，通过场景和拆解特征之间的关联关系构成场景的网状知识结构。本文的面向向量拆解结果的场景-特征的关系抽取算法的关键技术主要涉及两点：解释结果原始数据类型的转换和关系类型的定义。

解释结果原始数据类型的转换算法，主要进行将 IBD 模型处理后的输出结构的原始半结构化数据，即场景对应一个解释特征集合的数据结构，转换成场景-特征或特征-特征两两之间的关系。在关

系类型的定义上，本文将解释结果中存在的关系划分为两大类：（1）解释关系：场景和特征之间存在解释关系，关系由特征指向场景；（2）并列关系：同一类场景的解释特征之间存在两两并列关系，是一种双向关系。这两种关系中，解释关系是单向的，并列关系是双向的，可理解为图的无向边概念。

基于上述原始类型的转换和关系类型定义方法，即可构建出整个数据集的场景-特征关联关系网络，利用这个网络可构建可解释基拆解模型在输入数据集上的知识图谱。图 4 以“library”（图书馆）展示了进行关系抽取的结果，每个拆解特征与场景之间构建的解释关系。

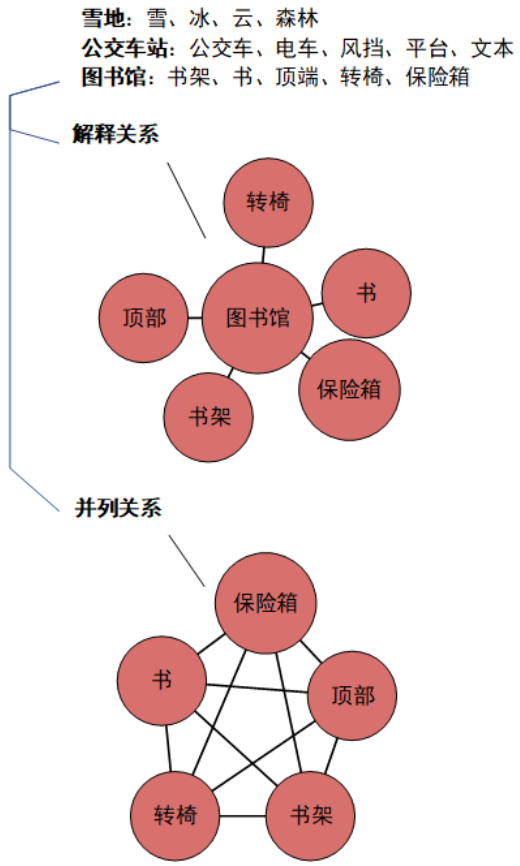

图 4 以“library”（图书馆）为例进行关系抽取的结果

然后针对不同的解释结果场景和特征间的不同关系，选择不同的抽取算法。针对第一种假设“场景和特征之间存在解释关系”的场景，本文对拆解结果进行解释关系抽取。在实现上，IBD 模型训练和测试分别使用了 Broden227 数据集和 places365 数据集，因此解释结果中类标签和特征标签不完全统一，交集和补集都非空。考虑到数据集标签不统一性，本文对解释结果关系抽取部分将不区分类标签和特征标签，因此解释关系中的实体包含类和特征实体，且解释关系图存在网状结构和层次关系。针对第二种假设“同一类场景的解释特征之间存在两两并列关系”，在神经网络中，可认为两个实体对应的拆解特征可共同表示某类结果，即特征之间可能存在共同协作关系，而不是单独发挥作用。本文对特征并列关系的抽取旨在探索是否存在特征之间共同作用、协同解释某一类别的现象。

基于上述设计思想，场景-特征关系提取算法的输入是场景和拆解特征的数组$Result$，以及数据集样本总数n，数组包含场景和 7 个拆解特征，以及每个特征在对应场景识别任务中的贡献度百分比。算法输出是场景-特征关系数组$RelationUpdown$和特征并列关系数组$RelationJoint$。算法首先从可解释基拆解模型的直接结果中，提取出特征指向场景的单向关系$RelationUpdown$，然后对整体样本中同类场景的所有拆解特征进行合并，得到每类场景与其所有可能的拆解特征集合的映射关系$relationJointGroup$，对每类场景的可能拆解特征集合中的贡献度最高的三个拆解特征之间构建特征之间的并列关系$RelationJoint$。伪代码如算法 1。

算法 1.场景-特征关系抽取算法

输入：场景和拆解特征数组$Result$，数据集样本总数n；

输出：场景-特征关系数组$RelationUpdown$，特征并列关系数组$RelationJoint$

1. $RelationUpdown$ = []
2. $RelationJoint$ = []
3. **FOR**i = 0 to n**DO**
   a) **FOR**j = 0 to 7 **DO**
      i. $RelationUpdown[i].start = Result.featurt_j$
      ii. $RelationUpdown[i].end = Result.label$
      END FOR
   END FOR
4. $RelationUpdown = RelationUpdown.groupby(RelationUpdown.start, RelationUpdown.end)$
5. $relationJointGroup = [(end, Set(RelationUpdown.start))forend, Set(RelationUpdown.start)\ in\ RelationUpdown.groupby(RelationUpdown.end))]$
6. **FOR**i = 0 to $RelationJointGroup.size$**DO**

a) $startSet = RelationJointGroup[1][:2]$

b) **FOR** $fea1$ $instartSet$ **DO**

i. **FOR** $fea2$ $instartSet$ **DO**

1. $RelationJoint.add(fea1, fea2)$

END FOR

END FOR

END FOR

7. **RETURN** $RelationUpdown, RelationJoint$

作为示例，图 5 展示了基于算法 1 的 IBD 模型对 resnet18 网络结构的解释结果，包含场景、特征以及解释关系和并列关系，使用 places365 的 validation 数据集进行知识图谱的构建，图中为所有关系中截取出的 1000 个关系样本。图 5 中节点表示场景和特征两种实体对象，关系包含场景-特征的解释关系和特征-特征的并列关系；图中黄色箭头表示场景-特征解释关系，表示特征指向场景，即解释概念指向被解释概念，蓝色箭头表示特征-特征并列关系，表示特征之间的双向关系，存在并列关系的特征很有可能共同表示某个或某几个场景，图 6 展示了节点“客厅”在 places365_val 数据集中的关联关系。图中节点表示场景和特征两种实体对象，关系包含场景-特征的解释关系和特征-特征的并列关系；图中黄色箭头表示场景-特征解释关系，表示特征指向场景，即解释概念指向被解释概念，蓝色箭头表示特征-特征并列关系，表示特征之间的双向关系，存在并列关系的特征很有可能共同表示某个或某几个场景，图中最能体现特征协同作用的是节点之间的蓝色边权重大小，权重越大，说明特征协同作用越明显。可以看出“客厅”概念的几个重要特征包括“椅子扶手”、“壁炉”和“阴影”等，且这三个特征之间构成并列关系环，检索“椅子扶手”、“壁炉”和“阴影”特征的解释关系，得到图 7 结果，图中有共同特征“阴影”的场景有客厅、大厅、卧室、楼层的隔层等，可用于探索不同概念的共同特征，关系深度为 2，除此之外，可通过两个概念的关系深度表示相似度，关系深度越小相似度越大，反之相似度越小。

本文将 IBD 解释模型的结果以知识图谱的形式进行存储，针对特征拆解结果的数据表现形式，定义了数据建模的相关概念，以图数据库的数据模块为基础框架，构建图数据库模块的节点、属性、关系、标签与特征拆解结果数据的映射关系，如表 1 所示，并以关系网络中”客厅“场景子图——图 6 为示例分别解释了这四个的概念。

4.3.2 基于 Jaccard 相似系数聚类的场景间距离和相似度计算

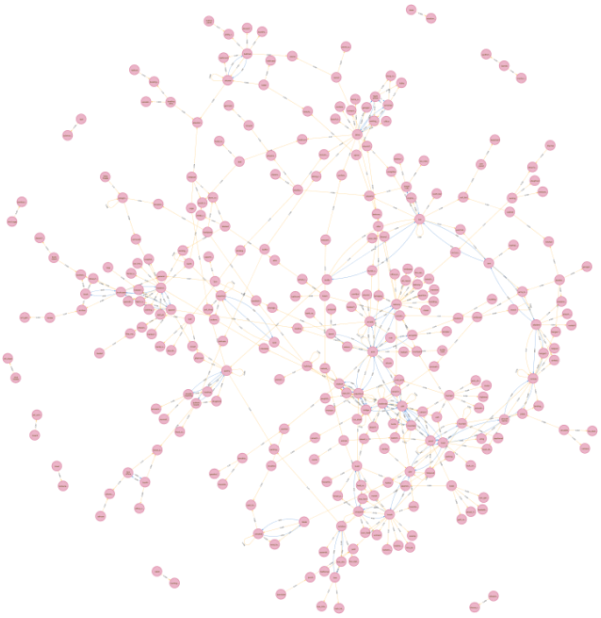

图 5 IBD 模型对 resnet18 网络结构的解释结果

传统计算场景间相似度的方法存在各种缺陷，如：词向量的相似度计算是基于训练好的词向量模型，针对不同的语言需要重新训练模型；词林计算相似度的前提是需要人工构建语料库的上下位关系等，比词向量模型的构建需要更大的投入。基于在不同语种下概念的不变性，本文首次提出一种对各种语言通用的概念相似度计算模型，可以避免因语言差异带来的重复工作，目前自然语言处理领域还没有类似模型。

基于上一节得到的场景-特征关系网络，本小节将从场景-特征的解释关系出发，提出一种基于场景-特征关联关系图的场景距离计算指标，可量化表示不同场景之间的距离，核心思想是用特征-场景的关系网络中场景共同特征数量表示两个场景间的相似度。理论上，两个场景的共同特征越多，表示两

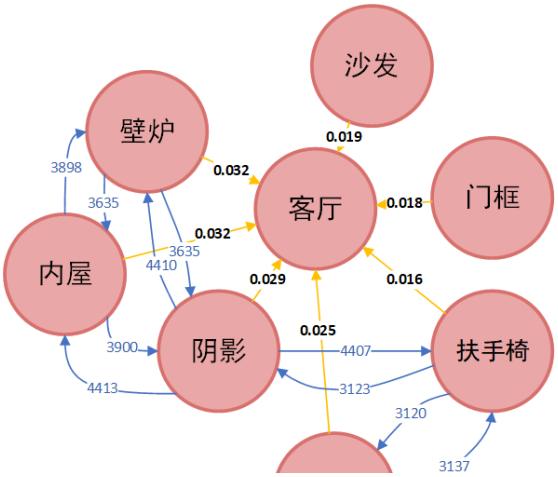

图 6 “客厅”概念的解释关系图

个场景相似度越高；反之，两个场景相似度越小。

因此，本文引入杰卡德相似系数（Jaccard Similarity Coefficient，本文简称 Jaccard 相似度）相似度计算的思想，进一步定义了场景间相似度的计算公式如公示（6）所示。

$$Sim(s_i, s_j) = \frac{|F(s_i) \cap F(s_j)|}{|F(s_i) \cup F(s_j)|} \#(6)$$

其中，$F(s_i)$表示场景s_i的拆解特征集合，$F(s_j)$表示场景s_j的拆解特征集合，$Sim(s_i, s_j)$为场景s_i和场景s_j之间的相似度。Jaccard 相似度适用于符号度量的个体间相似度，而本文 IBD 的解释结果构成为特征和场景，场景标签的解释为一系列特征标签，可视为一种符号度量。因为场景解释为一系列离散的特征标签，无法将场景的特征转化成数值表示形式，所以可以从是否具有相同特征来衡量场景之间

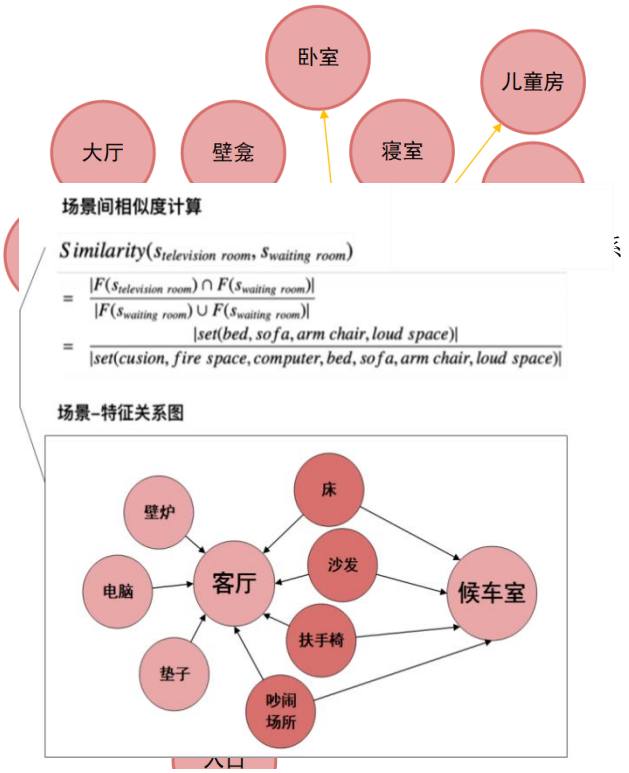

图 8　场景-特征关系图

的差异。图 8 是采用本文的场景间相似度算法得到的相似度计算结果示意图，以客厅和等候室两个场景为例，图中节点表示客厅和等候室这两个场景以及可解释基拆解模型对它们的 7 个拆解特征。图中，场景“living roomg”（客厅）和“waiting room”（等候室）的相似度计算示意图，图中是可解释基拆解模型对客厅和等候室的解释结果，其中客厅提取了 7 个特征，分别是：窗帘、壁炉、计算机、床、扶手椅、沙发和大空间，等候室提取了 4 个特征，分别是：床、扶手椅、沙发和大空间，则客厅和等候室的相似度是二者的共同特征个数(这里是 4 个）除以二者总特征数集合大小（这里是 7），所以客厅和等候室的相似度约为0.57。

与相似度计算公式相对应，本文定义场景间距离计算公式如公示（7）所示。

$$D(s_i, s_j) = 1 - Sim(s_i, s_j) = 1 - \frac{|F(s_i) \cap F(s_j)|}{|F(s_i) \cup F(s_j)|} \#(7)$$

本文提出了基于聚类的场景距离算法，该算法中对数据集中的每个类，使用距离计算公式计算出类间距$d_{s_i,s_j} \in M \subset R^D$，其中$M$为二维距离矩阵，$R^D$表示维度为D的数据集。例如，当用 places365_val 数据集进行聚类时，类别D=365。

场景间距离矩阵算法假定最远距离为 1，最近距离为 0，即为解释特征集合完全重合的两个场景。算法输入为场景-特征上下位置解释关系数组$Explanation$和需要计算距离的场景数n，输出为场景间相似度矩阵M。关系数组$Explanation$的元素是一个字典，字典的键是场景标签，值是该场景的解释特征集合，n为大于 0 且不超过关系数组$Explanation$最大长度的整数。首先使用距离计算公式计算出类间距离矩阵，是一个大小为$365 * 365$的二维矩阵M；使用 networkx 工具对距离矩阵M构图，最终得到 places365_val 数据集的聚类结果可视化效果（图 9）。本节构建的聚类效果图为全连接图，图中红色节点表示不同的类，边表示类与类之间的邻接关系，点之间的距离是依据距离计算该式的缩放效果。伪代码如算法 2。

算法 2.基于聚类的场景间距离矩阵算法

输入：$Explanation$场-特征上下位置解释关系数组，n需要计算距离的场景数

输出：场景间驱离矩阵M

1. $M = 1$
2. **FOR**i = 0 to n**DO**
   a) **FOR**j = 0 to n**DO**
      i. $M[i][j] = 1 - len(CACHE[i]CACHE[j])/len(CACHE[i]|CACHE[j])$
      END FOR
   END FOR
3. **RETURN**M

用场景间距离计算公式在 places365_val 数据集上的聚类效果如图 9 所示。本节从 places365_val 数据集抽取 100 个类进行聚类可视化，图中红色节点表示不同的类，边表示类与类之间的邻接关系，点之间的距离是依据距离计算公式的同比缩放效果，图中两点距离越近表示两个类相似度越高，距离越远表示两个类相似度越低。基于 Jaccard 相似系数的场景间相似度度量方法优点是计算复杂度向度较低，为$O(n^2)$，且这种相似度计算方法是严格依赖于特征拆解结果的，可以直接反映出深度神经网络经过迭代训练后对数据集不同场景的理解。

4.4 判别性特征提取

4.4.1 判别性特征定义

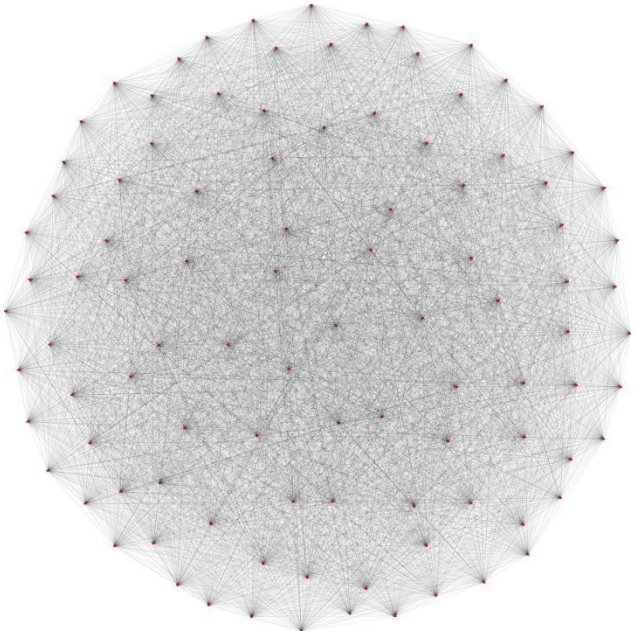

图 9 用场景间距离计算公式 7 在 places365_val 数据集上的聚类效果图

现有的 IBD 解释模型的场景特征拆解结果，均以百分比的形式表示不同特征对场景样本识别的贡献度，其中，百分比可以抽象成该特征对目标场景的重要性指数，其与特征对目标场景的重要性成正比。对于两个相似的场景，其解释特征重合率可能很高，这种情况下无法通过贡献率最高的特征来区别两个相似场景。因此本文探索是否存在一个或多个能够较准确地区分不同场景的特征，本文将这种特征成为“判别性特征”。

通过分别计算特征在所有场景和特定场景的解释中出现的频率可以得到该特征对全局场景和特定场景的重要性。理论上，结合以上两个度量指标，通过计算该特征对全局场景和特定场景的重要性的权衡结果来度量判别性特征的判别能力。

在神经网络解释模型的拆解结果中，特征和场景标签可类比语料库，因此，本文提出将 TF-IDF 的思想类比到神经网络可解释性中判别性特征的提取，找到一种合理的从 IBD 解释模型特征拆解结果到自然语言处理领域基本概念的映射关系。

4.4.2 理论基础

TF-IDF （ Term Frequency–Inverse Document Frequency）的核心思想是：用元素在本类中的出现频率表示类内重要性，用元素在所有样本中的出现频率表示类普遍重要性，最后通过衡量上述两个重要性指标的出元素对本类的判别性能。TF-IDF 主要关心两大指标：类内重要性和普遍重要性。本文将借助 TF-IDF 的思想，提出一种新的特征提取方法，用以评估一个特征对于数据集中一类场景的重要性。

4.4.3 概念映射关系构建

表 2 列出了 NLP（Natural Language Processing）的基本概念与 IBD 解释模型的特征拆解结果之间的映射关系，本文将 IBD 模型的拆解特征视为语料库中的单词f，一类样本的所有特征拆解结果集合视为语料库中的文本S，而 IBD 模型测试用到的数据集（本节使用 places365 的校验数据集）作为语料库R。则 TF-IDF 计算过程中，拆解特征在所有样本在拆解结果中的占比即为词频$TF_{f,S}$，计算结果为拆解特征f的普遍重要性和对场景S的重要性的综合权衡量化表示的结果。基于表 2 中的概念映射关系，本小节提出适用于场景数据集的场景 TF-IDF 指标计算方法。

表 1 图数据库模块建模概念定义

数据模块	定义	示例
节点	特征/类型实体对象	图 6 中的场景“客厅”，以及“椅子扶手”、“壁炉”和“阴影”等特征
属性	节点入度/出度，关系的同类样本平均权重/同类样本数	图 6 中存在特征“椅子扶手”
关系	解释/并列关系	图 6 中场景和特征之间存在的有向边，如“椅子扶手”和”客厅“节点之间的边
标签	节点类/特征标	图 6 中节点上标注的名称，如

	签，关系解释/并列类型标签	“客厅”节点对应图中“living room”标签

首先计算特征出现频率，即某个特征在所有样本的拆解结果中的出现频率，表示特征f在场景S的拆解结果中出现的频率，$n_{f,S}$为特征f在场景S的拆解结果中出现的次数，$\sum_k n_{k,S}$表示场景S的所有拆解特征个数，$TF_{f,S}$为特征f在场景S的解释结果中的出现频率。特征对场景的频率，可以理解为对拆解特征数量的归一化，特征对场景的频率越高，说明这个特征对其解释的场景重要性越大，这个指标可类比可解释基拆解模型的特征贡献度指标意义。

$$TF_{f,S} = \frac{n_{f,S}}{\sum_k n_{k,j}} \#(8)$$

拆解特征的普遍重要性指的是在语料库中包含该词的文档数。$|R|$为语料库中的总文档数，t_f为某个特征，R_S为一类场景，$\left|\{S: t_f \in R_S\}\right|$表示数据集中包含特征$t_f$的场景数，$IDF_f$为特征$t_f$对数据集中所有场景的普遍重要性。可以通过数据集中的总场景类别数除以包含该特征的场景总数，再取对数得到。如果包含该特征向量的场景数越少，即该特征的普遍重要性越小，IDF_f值越大。反之包含该特征向量的场景数越多，即该特征的普遍重要性越大，IDF_f值越小。IDF 是度量特征普遍重要性的量化指标。

$$IDF_f = lg\frac{|R|}{\left|\{S: t_f \in R_S\}\right|} \#(9)$$

最终 TF-IDF 由特征出现频率 TF 和普遍重要性 IDF 两者相乘得到。$TF_{f,S}$为特征f在场景S的拆解特征中的出现频率，IDF_f为特征t_f在数据集中的普遍重要性指标，得到特征f对于场景S的 TF-IDF 指标$TFIDF_{f,S}$。如果一类场景的解释结果中的高频特征，在整个数据集中的 IDF 指数也很高，则认为该特征对于给定场景有很好的判别能力。

$$TFIDF_{f,S} = TF_{f,S} \times IDF_f \#(10)$$

4.4.4 判别性特征提取算法实现

算法的输入是场景-特征的关系数组$Relation$，数组的元素为二元组的形式(fea, s)，其中s为一类场景，fea为场景s的拆解特征之一，用二元组(fea, s)表示存在由特征fea指向场景s的解释关系；同时，对输入的每对解释关系二元组，分别对应一个解释关系对解释场景的贡献度$Weight$，表示这样的解释关系在所有解释结果（解释关系二元组）中的占比；第三个输入是所有解释结果中，包含拆解特征fea的场景类别数，以数组$Count$形式输入。输出是每个解释关系的 TF-IDF 指标数组$TFIDF$。伪代码如算法 3。

算法 3.判别性特征提取算法

输入：场景-特征解释关系数组$Relation$, 场景-特征解释关系贡献度数组$Weight$, 场景-特征解释关系出现频数数组$Count$

输出：场景-特征的TF − IDF指标数组$TFIDF$

1. $TF = Weight(fea, s)/Count(fea, s)$
2. **FOR**(fea, s) in $Relation$ **DO**
   a) $CF[fea] = CF[fea] + 1$
   END FOR
3. $IDF = log10(TF\ .size/CF\ ))$
4. TFIDF = TF ∗ IDF
5. **RETURN**$TFIDF$

4.4.5 判别性特征提取结果

本节测试使用本部分知识图谱内的所有数据，即 places365_val 数据集，resnet18 训练迭代 14 个 epoch，计算 TF-IDF 指标，表 3 列出 Top10 结果，

表 2 TF-IDF 符号-概念映射表

符号	NLP 概念	IBD 概念
f	单词	可解释基概念（拆解特征）
S	文本	一类场景的特征拆解结果集合
R	语料库	场景数据集
$\|R\|$	语料库文件总数	数据集中的总场景类别数
$TF_{f,S}$	词频	拆解特征在所有样本的拆解结果占比
$\|S: t\_f \in R\_S\}\|$	包含单词的文件数	包含拆解特征的场景类别数
IDF_f	词的普遍重要性	拆解特征的普遍重要性
$TFIDF_{f,S}$	词 f 对文档 S 的 TFIDF 指标	特征 f 对于场景 S 的 TFIDF 指标

表 3 TFIDF 指标测试结果 Top10

序号 Top-n	TFIDF	场景	拆解特征
1	0.3995	skyscraper	**skyscraper**, tower, chimney, monument, cloud
2	0.3442	waterfall	**waterfall**, cliff, rock, pond, fog bank
3	0.3282	library	**bookcase**, book, top, swivel chair, vault
4	0.3279	cockpit	**controls**, instrument panel, airplane, steering wheel, console table
5	0.3277	basketball court	**court**, tennis court, table tennis, flag, beak
6	0.3227	carrousel	**carousel**, organ, foot, chain wheel, playground
7	0.3119	snowfield	**snow**, ice, cloud, clouds, forest
8	0.3067	amusement arcade	**arcade machine**, slot machine, screen, table game, table tennis
9	0.3034	escalator	**escalator**, conveyer belt, riser, bowling alley, shops
10	0.2971	bus station	**bus**, autobus, windshield, platform, text

其中拆解特征加粗部分为根据 TF-IDF 指标筛选得到的最具判别性特征。

在判别性特征提取时，存在类标签和判别性特征相同的情况，原因是特征训练集 Broden 数据集包含 places365 数据集的部分标签，即类标签集合和特征标签集合存在部分交集，在表中表现为将第三列的类标签作为第四列加粗判别性特征的情况。这样类标签和判别性特征重叠的情况可以反应出场景数据集的 365 种场景标签是确切的场景样本，而 IBD 模型使用的特征向量的训练集标签粒度不够细，标签粒度不统一导致某些场景样本的特征拆解结果难以具体定位。但除此之外的其他拆解特征（TF-IDF Top-5 特征)可作为场景的有效拆解特征，如 skyscraper（摩天大楼）的 TF-IDF Top-5 特征中 tower（建筑物的塔形部分）、chimney（烟囱）、monument（纪念碑）、cloud（云），可以看出 TF-IDF 指标更高的特征更具有区别其他标签的效果，可以看出 TF-IDF 指标更高的特征更具有区别其他标签的效果，可以看出 TF-IDF 指标更高的特征更具有区别其他标签的效果，提取出的判别性特征主要刻画了大楼的轮廓特征：高耸、长方体等，而 TF-IDF 分数靠后的特征 cloud 则关注点转移到了场景的非主体部分：天空。因此以 TF-IDF 指标为依据进行分析更能抓住特定场景下能区别于其他场景的特征，而不是仅仅考量神经网络对特定某个场景的特征捕获程度。

图 10 统计了 places365_val 数据集场景标签的样本分布情况，图 11 统计了 places365_val 数据集每一类场景的 TF-IDF 指标 Top-5 的均值。

5 测试与验证

测试是对模型正确性的客观验证，对神经网络可解释性的测试应该至少覆盖两个角度：人类置信度和保真度。人类置信度即解释结果是否符合现实中人类对事物的认知，解释是否合理；保真度即解释结果是否遵从神经网络本身，是否正确表达神经网络的决策依据/过程。

在上述两种测试角度中，现有的研究对可解释基拆解模型的测试仅覆盖前者——人类置信度的度量，而没有对后者——解释模型的保真度做量化测试，因此模型存在测试维度不全面的问题。本文则从人类置信度和保真度提出了新的测试方法，并对方法进行了测试。

5.1 基于哈德玛积的保真度测试

本文对可解释基拆解模型结果中，拆解特征在像素级的定位能力进行量化测试，其定位表现形式为热谱图，通过测试热谱图来探索可解释基拆解模型的特征拆解结果在像素级别的定位能力。

受 Aditya 等人提出的热谱图的识别主体定位能力测试方案[13]的启发，本部分将使用可解释基拆解模型的多特征热谱图定位能力表示可解释基拆解模型的模型保真度。不同于使用类热谱图的 Grad-CAM++模型[13]，本文模型结果为多特征热谱图，每个拆解特征对应一个热谱图，可以认为本文模型结果对一个场景的拆解特征热谱图之和与 Grad-CAM++模型[13]的类热谱图是同类概念，其表示的意义均可认为表示一类场景的综合特征。

首先需要完成多特征热谱图的融合工作，即得到场景综合特征热谱图，以便后续量化测试的展开和分析。用亮度表示深度神经网络对该输入图像的识别区域，越亮的区域表示深度神经网络在这个区域的激活程度较高，反之则表示深度神经网络激活程度较低。

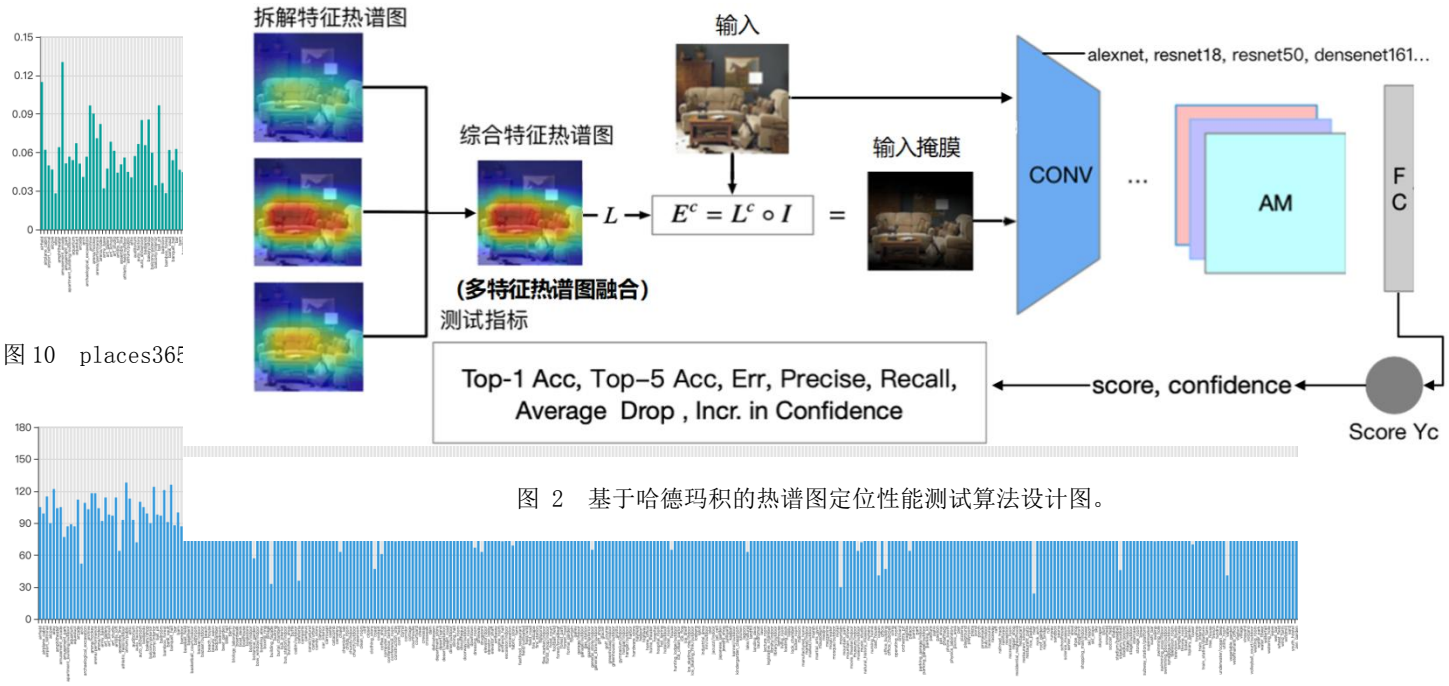

图 10 places365

图 2 基于哈德玛积的热谱图定位性能测试算法设计图。

图 1 places365_validation 数据集每一类场景的 TF-IDF 指标 Top-5 的均值。

本部分使用目标识别任务中对每类场景的置信度变化率体现这一性能，置信度提升/降低的变化率大说明由综合热谱图提取重点识别区域的能力不稳定，相反，置信度提升/降低的变化率小说明热谱图提取重点识别区域的能力稳定。

与[13]中对类的热谱图的测试方案类似，本文提出一种基于哈德玛积的热谱图定位性能测试算法，算法测试设计图如图 12 所示。图中左边有两个输入数据集——原图数据集和 mask 数据集，其中 mask 数据集为可解释基拆解结果中的拆解特征热谱图 Top-3 结果的融合结果与原图作哈德玛积运算，得到的 mask 数据集，将这两个对比数据集分别输入到深度卷积神经网络中，对比 Top-1 准确率，Top-5 准确率，错误率，精确度，召回率，特征融合数据集置信率下降样本集合的平均下降率和置信率提升样本集合的平均提升率这 7 个场景识别任务的模型性能度量指标。

5.1.1 基于哈德玛积的特征融合图谱获取算法

对可解释基拆解模型结果中每个场景的贡献度指标 Top-3 的拆解特征，取三者激活域的最大激活值，获得类c的激活图谱L^c如公示（11）所示。

$$L^c(x,y) = \max_{i\in[1,3]} L_i^c(x,y) \#(11)$$

其中$L_i^c(x,y)$是类c其中一个输入样例的特征i激活图谱在(x,y)位置的激活值，$L^c(x,y)$为类c其中一个输入样例的特征融合激活图谱在(x,y)位置的激活值，本文取重要性 Top-3 特征在(x,y)位置的最大值。对于获得L^c为类c某样例下的特征融合激活图谱，本文采用哈德玛积（Hadamard product）公式获得类c的特征融合图谱数据集E^c。

$$E^c = L^c \circ I \#(12)$$

其中，L^c为类c的激活图谱，I为输入的原图，E^c为类c的特征融合图谱，$\circ$为哈德玛积运算。图 13 展示了原图、特征热谱图和特征融合图谱的转换效果。

特征融合数据集用于与原图数据集做对比测试，分别测试 alexnet、resnet18 等不同 CNN 网络结构在这两个对比数据集上的识别准确度，主要对比的指标有：Top-1 准确率、Top-5 准确率、错误率、精确度、召回率、特征融合数据集置信率下降样本集合的 Average Drop 和置信率提升样本集合的

Confidence Incr.。由于不同网络结构的网络深度差异，在有限的计算资源上，本节实验对 alexnet、resnet18、resnet50、densenet161 分别使用的训练参数 batch size 为 16，16，32 和 8。

5.1.2 测试数据集

实验在 Places365 数据集的校验集上进行验证，Places365 是 MIT 整理的场景图像数据集，包含 1 千万张图片，其校验集包含 36500 个样本，365 类，每类 100 个样本，可用于以场景和环境为识别主体的视觉认知任务。因为在 places365 场景识别数据集中，包含更多物体，这些物体可作为场景特征，即场景作为分类结果，场景中包含的物体作为拆解结果中的概念特征向量。这样的拆解结果更具有可解释性，且将场景中的物体作为场景识别依据，更方便后续人工评判 IBD 模型的拆解向量结果的可解释性。

5.1.3 保真度测试结果分析

本节使用可解释基拆解模型对 alexnet、resnet18、resnet50、densenet161 四种网络结构分别训练，得到解释结果，并通过本节提出的特征融合图谱获取算法得到测试对比数据集——特征融合数据集。

本节分别对 4 个不同网络结构在 2 个对比数据集——原图数据集（places365_val 数据集）和特征融合数据集上进行场景识别任务性能测试，测试结果如表 4 和表 5 和表 6 所示。对同一网络结构，相较于使用原图数据集，特征融合数据集 Top-1 准确率、Top-5 准确率、精确度和召回率均呈现约 20%比例的下降，错误率呈现约 20%比例的上升。

alexnet、resnet18、resnet50、densenet161 四个网络结构在 places365_val 数据集上的准确率对比如图 14 所示。

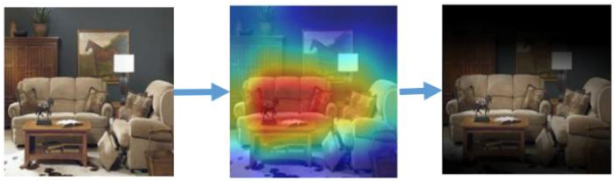

图 3 原图、特征热谱图和特征融合图谱的转换效果。

表 4 中从网络结构的维度分别测试了 5 项模型性能指标，这些数据对比结果整理到表 5 和表 6 中。Top-1 准确度、Top-5 准确率、精确度和召回率这四项指标的下降率越高，错误率的提升率越高，表示可解释基拆解模型对该网络结构的目标识别区域定位越不准确，保真度越低。

从表 5 和表 6 中数据可以看出上述 4 个不同网络结构 Top-1 准确率平均降低幅度约 25.08%，对置信率提升样本集合的 Confidence Incr.比置信率降低样本集合的 Confidence Average Drop 高 1.83%，即可解释基拆解模型提取的深度神经网络在原图中像素级的识别区域对模型置信度提升了 1.83%。

从表 5 和表 6 中数据可以看出上述 4 个不同网络结构 Top-1 准确率平均降低幅度约 25.08%，对置信率提升样本集合的 Confidence Incr.比置信率降低样本集合的 Confidence Average Drop 高 1.83%，即可解释基拆解模型提取的深度神经网络在原图中像素级的识别区域对模型置信度提升了 1.83%。

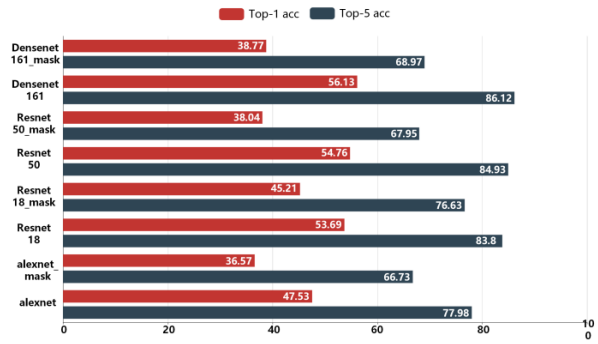

图 4 特征融合数据集在 places365 数据集上的准确率对比结果

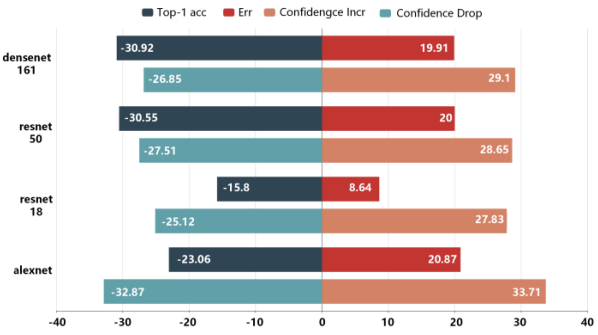

图 5 特征融合数据集在 places365 数据集上的 Top-1 准确率、错误率、特征融合数据集置信率下降样本集合的平均下降率和置信率提升样本集合的平均提升率的变化对比结果。

resnet18 的 Top-1 准确率、Top-5 准确率、精确度、召回率下降幅度最小，分别为 15.8%、8.64%、18.33%、10.27%、15.8%，
说明使用由特征热谱图融合生成的特征融合图作为输入，对 resnet18 的场景识别性能影响最小，
可以认为原图中真正具有识别能力的部分与特征融合热谱图中激活值较高的像素区域基本吻合。

resnet18 的置信率降低样本集合的 Average Drop 指标最低，为 25.12%，综合前面所述 Top-1 准确率、
Top-5 准确率、精确度、召回率下降幅度最小的现象，可以推测可解释基拆解模型作用于 resnet18 网络结构的解释结果保真度最高。

alexnet 的置信率提升样本集合的 Confidence Incr.指标最高，为 33.71%，即在 alexnet、resnet18、resnet50、densenet161 四种不同网络结构中，可解释基拆解模型对 alexnet 的置信率提升样本的提升率最高，反映出对 alexnet 网络结构解释结果的热谱图对物体识别的定位能力最强。但是 alexnet 的置信率降低样本集合的 Average Drop 指标也最高，为 32.87%，反映出对 alexnet 网络结构的解释结果中热谱图定位能力不稳定。

表 5 和表 6 的数据的变化率如图 15 所示，图 15 展示了特征融合数据集在 places365 数据集上的 Top-1 准确率、错误率、特征融合数据集置信率下降样本集合的平均下降率和置信率提升样本集合的平均提升率的变化对比结果，纵轴是不同的网络结构，横轴表示上述四个指标的变化率（%），正数表示在对应指标测试中，使用特征融合数据集比使用原图数据集有所提高，负数表示使用特征融合数据集比使用原图数据集有所下降。

从本部分测试的对比实验可推测，可解释基拆解模型对 alexnet、resnet18、resnet50、densenet161 这 4 种不同网络结构的解释能力不同，说明可解释基拆解模型的解释能力与网络结构、网络深度有很大关联。

5.2 基于TF-IDF差异的解释结果人类置信度测试

除了模型的保真度测试，本文还覆盖了解释结果的人类置信度测试，测试的目的是度量 IBD 模型对场景的解释结果与人类理解的差异程度。为了量化表示这一差异，本文对同一网络结构的测试结果使用这些判别性特征 TF-IDF 值的 MSE（Mean Square Error）指标度量模型解释与人类理解之间的差异。

表 4 特征融合数据集在 places365 数据集上的测试结果 1

网络结构-数据集	Top-1 acc	Top-5 acc	Err	Precise	Recall
alexnet-places365_val (base)	0.475378	0.779791	0.002875	0.462651	0.475457
alexnet-places365_val_mask	0.365772	0.667262	0.003475	0.395429	0.365754
resnet18-places365_val (base)	0.536932	0.838000	0.002537	0.532317	0.536932
resnet18-places365_val_mask	0.452110	0.766301	0.003002	0.477668	0.452110
resnet50-places365_val (base)	0.547649	0.849326	0.002479	0.542619	0.547636
resnet50-places365_val_mask	0.380350	0.679489	0.003395	0.460973	0.380342
densenet161-places365_val (base)	**0.561294**	**0.861218**	**0.002404**	**0.556521**	**0.561282**
densenet161-places365_val_mask	0.387747	0.689709	0.003355	0.477039	0.387739

表 5 特征融合数据集在 places365 数据集上的测试结果 2

网络结构-数据集	Top-1 acc Drop%	Top-5 acc Drop%	Err Incr.%	Precise Drop%	Recall Drop%
alexnet-places365_val_mask	23.06	14.43	20.87	14.53	23.07
resnet18-places365_val_mask	**15.8**	**8.64**	**18.33**	**10.27**	**15.8**
resnet50-places365_val_mask	30.55	20	36.95	15.05	30.55
densenet161-places365_val_mask	30.92	19.91	39.56	14.28	30.92
Average	25.08	15.75	28.93	13.53	25.09

表 6 特征融合数据集在 places365 数据集上的测试结果 3

网络结构-数据集	Average Drop% (Lower is better)	Confidence Incr.% (Higher is better)
alexnet-places365_val_mask	**32.87**	**33.71**
resnet18-places365_val_mask	**25.12**	**27.83**
resnet50-places365_val_mask	27.51	28.65
densenet161-places365_val_mask	26.85	29.10
Average	28.09	29.82

表 7 神经网络结构的识别准确率

网络结构	alexnet	resnet18	resnet50	densenet161
Top-1 acc %	47.54%	53.69%	54.76%	56.13%
Top-5 acc %	77.98%	83.80%	84.93%	86.12%

5.2.1 场景-特征数据来源

人工评测以调查问卷的形式展开，测试人员一共有 5 人，涵盖了解和不了解神经网络的人员。由于 places365 数据集的标签均为英文格式，为了方便评测展开，本文使用百度翻译接口对 places365 数据集的英文标签直译成中文格式。本文的测试数据由 places365 数据集中抽取场景识别任务中平均置信度最高的 20 类场景组成。分别对 alexnet、resnet18、resnet50、densenet161 四种网络结构展开判别性特征准确度的测试。

测试重点是 IBD 解释结果的 TF-IDF 维度的判别性特征提取的合理性。本文对不同的网络结构分别展开人工测试，测试形式为调查问卷，内容以类标签维度展开，以单项选择的形式，需要测试人员从 IBD 特征拆解结果的 TF-IDF 指标 Top5 的特征中选择一项最具有判别性的特征。选项中混入的 TF-IDF 指标最高的特征，记为c_{us}，测试人员从 IBD 拆解结果中选出一个认为最合理的判别性特征，记为c_{human}。在所有结果中，本文假定投票占比最高的选项为人工评测最合理的判别性特征c_{human}。对于最具判别性特征c_{us}和人工选取的最具判别性特征c_{human}之间的距离，本文指定一种距离计算方法，目的是得到两个特征之间的偏差。利用计算特征i_1和特征i_2对类j的 TF-IDF 值，计算两个特征 TF-IDF 的 MSE 指标。

表 8 神经网络结构的人工评测与 IBD 解释结果重合率

网络结构	alexnet	resnet18	resnet50	densenet161
人工评测与 IBD 解释结果重合率%	54.00	**60.83**	44.00	**23.00**

表 9 神经网络结构的人工评测与 IBD 解释结果偏差

网络结构	alexnet	resnet18	resnet50	densenet161
人工评测与 IBD 解释结果偏差 MSE	0.000132	**0.000413**	0.000002	0.000002

5.2.2 判别性特征提取准确度测试结果分析

表 7 为本次测试使用的不同网络结构：alexnet、resnet18、resnet50、densenet161，在 Broden 数据集上的识别准确率指标 Top-1 准确率和 Top-5 准确率。表 8 为本次测试中人工测试结果，指标百分比表示神经网络结构的人工评测与 IBD 解释结果重合率，展示了 alexnet、resnet18、resnet50、densenet161 四个网络结构的人工评测与可解释基拆解模型的判别性特征重合率百分比。

为了进一步探究人工评测与基于 TF-IDF 指标的判别性特征提取结果的差异大小，本文使用 TF-IDF 指标量化表示人工评测中的非重合结果（以下简称：非重合结果）与基于 TF-IDF 指标提取的判别性特征之间的偏差。算法设定二元组$(\mathrm{case}^{c}_{\mathrm{IBD}}, \mathrm{case}^{c}_{\mathrm{bad}})$表示对类标签为$c$的最具判别性特征提取结果，其中$\mathrm{case}^{c}_{\mathrm{IBD}}$为基于 IBD 解释模型结果提取的 TF-IDF 指标为导向的最具判别性特征提取结果，$\mathrm{case}^{c}_{\mathrm{bad}}$为人工评测中出现的非重合结果。通过计算每一类的 TF-IDF 指标在拆解特征维度的方差，量化表示非重合结果的偏差程度。

表 9 为本次人工测试结果与 IBD 解释结果偏差，

使用 MSE 指标量化表示 IBD 解释结果在 TF-IDF 维度的偏差。从表中数据可以看出人工评测重合率最高的 resnet18 网络结构的非重合结果偏差最高，MSE 指标为 0.000413，相反，人工评测重合率最低的 densenet161 网络结构的非重合结果偏差最低，MSE 指标为 0.000002。

综合以上结果，可以推测，可解释基拆解模型对 resnet18 网络结构的解释较为精准，且判别性特征与非判别性特征的 TF-IDF 指数差异较大，即 resnet18 网络对识别目标的局部特征提取针对性较强。相反，可解释基拆解模型对 densenet161 网络结构的解释较不精准，且判别性特征与非判别性特征的 TF-IDF 指数差异很小，可推测可解释基拆解模型对 densenet161 网络结构的解释较为模糊，场景的解释特征判别性指向不明显，反映出 densenet161 网络对识别目标的各部分特征提取较均衡。

从准确度测试结果来看，在判别行特征提取的重合率和整体偏差度量指标 MSE 这两个维度，人工评测与 IBD 解释结果的重合度与非重合结果对应 TF-IDF 值的 MSE 指标呈正比，即用 TF-IDF 提取的判别性特征与人类理解的重合度越高的网络结构，其非重合结果的 TF-IDF 偏差越低；重合度越低的网络结构，其非重合结果的 TF-IDF 偏差越高。可从这一现象推测不同网络结构的训练学习方式也不同，与人工评测重合度高且非重合结果的 TF-IDF 偏差高的网络结构，其对数据集中的场景特征学习针对性较强，即倾向于学习某一个或几个特征作为场景的识别主要特征；而与人工评测重合度低且非重合结果的 TF-IDF 偏差低的网络结构，其对数据集中的场景特征学习针对性较弱，即倾向于学习某几个特征的综合结果作为场景识别的主要特征。

6 总结与展望

可解释基和知识图谱是研究深度神经网络可解释性的两项前沿技术，本文针对现有深度学习网络可视化研究存在的可解释基模型结果输出缺乏基于图谱进行解释和缺乏量化测试等问题，提出了一种基于可降解基拆解和知识图谱的深度神经网络内部机制可视化方法，据我们的调研，我们的研究是国内外第一次融合了可解释基和知识图谱的工作，量化实验证明了本文方法的优异效果。

本文的工作主要分为两大点：第一，用知识图谱的形式对可解释基拆解模型的结果构建了更全面的解释；第二，对可解释基拆解模型的模型保真度和人类置信度这两个方面进行全面的测试和分析。本文首先对可解释基拆解模型的解释结果进行后处理，利用场景-特征的关联关系构建了深度神经网络分类模型的知识图谱；并提出一种基于 Jaccard 相似系数的场景间相似度计算方法，在同一网络结构维度上可以对任意两个场景相似度进行度量，算法严格依赖于特征拆解结果，可以直接反映出深度神经网络经过迭代训练后对数据集不同场景的理解，且是基于训练集和网络结构两个维度上的相似度计算方法，这种相似度计算以图像的形式理解人类社会的概念，优势是可以跨越语言的差异，以绝对的事物概念理解不同场景，另外，这种相似度表示方式与人类理解事物的方式接近，即这种相似度计算公式容易加入人类的其他外部知识，可扩展性强，算法的缺点是无法量化表示完全不相关的两个场景间的相似度差异，仅适用有一定共同点的两个场景间相似度的计算；本文还提出一种基于 TF-IDF 的场景判别性特征的提取方法，用人工评测的方式对判别性特征提取的准确度进行量化测试，测试目的在于探索不同模型对识别目标的理解与人类理解方式的差异。除此之外，本文还对可解释基拆解模型的保真度性能进行量化测试，通过对模型结果的多特征热谱图的定位能力，抽象表示可解释基拆解模型的保真度，将多特征热谱图结合成一个综合特征热谱图，用哈德玛积将综合特征热谱图与原图再融合，得到特征融合图谱，用特征融合图谱的亮度标记深度神经网络分类模型对识别目标的定位像素区域，进而通过对比输入原图和特征融合图谱对模型分类能力的偏差大小衡量热谱图的目标定位能力，即可解释基拆解模型的保真度。

本文的未来工作包括：本文对可解释基拆解模型的解释结果构建了深度神经网络分类模型的知识图谱，知识图谱严格依赖于训练模型使用的数据集，因此灵活性很强，可以通过替换不同类型场景的数据集获得该深度神经网络模型对不同类型数据集的知识构建结果；除此之外，以知识图谱的形式抽象表示深度神经网络模型的内部知识构建，形式与人类构建的外部知识图谱无差异，因此用知识图谱的形式抽象深度神经网络模型的内部知识，较易与外部知识图谱融合，本质上来说，是人类理解与机器理解的结合，未来可研究人类与机器理解相

结合的知识结构变化。此外，基于 TF-IDF 的场景判别性特征的提取方法应用场景广泛，未来可利用场景的判别性特征优化场景文本的关键词提取过程，以及可以通过更换训练数据集帮助特定场景类型的识别准确性提高，如可用于社区安防，训练集使用犯罪场景集，通过可解释基拆解模型对犯罪场景识别预训练模型进行解释，得到犯罪场景的判别性特征，用这些判别性特征标记的数据集对原犯罪场景数据集进行扩充，对犯罪场景识别模型再训练，可帮助模型抓取更多相关场景的关键特征，提高模型识别准确率。下一步的工作计划：包括由于知识图谱依赖于训练模型使用的数据集，灵活性很强，本文将尝试扩展不型场景的数据集，在更广泛的场景（比如安防等场景集）获得该深度神经网络模型对不同类型数据集的知识构建，这将进一步扩展本文方法的研究意义和应用价值。

参 考 文 献

[1] He K, Zhang X, Ren S, et al. Deep Residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).las Vegas, USA, 2016:770-778

[2] A. Graves, A. Mohamed and G. Hinton, Speech recognition with deep recurrent neural networks// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.Vancouver, Canada, 2013:6645-6649

[3] Zhou B, Sun Y, Bau D, et al. Interpretable basis decomposition for visual explanation//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018:119-134

[4] Zhang Q, Yang Y, Ma H, et al. Interpreting cnns via decision trees//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6261-6270

[5] Y. Liu, Y. Xie and A. Srivastava. Neural Trojans//Proceedings of the IEEE International Conference on Computer Design (ICCD).Boston, MA, 2017:45-48

[6] Ribeiro M T, Singh S, Guestrin C.. " Why should I trust you?" Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[7] Zilke J R, Menc á E L, Janssen F. Deepred–rule extraction from deep neural networks// Proceedings of the International Conference on Discovery Science. Bari, Italy, 2016: 457-473

[8] Zeiler M D, Taylor G W, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning// Proceedings of the International Conference on Computer Vision.Barcelona, Spain, 2011: 2018-2025

[9] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks// Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 818-833

[10] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences//Proceedings of the 34th International Conference on Machine Learning.Sydney, Australia, 2017: 3145-3153

[11] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,USA,2016: 2921-2929

[12] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision. 2020, 128(2): 336-359

[13] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks// Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA. 2018: 839-847

[14] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2015:3-8

[15] Oztireli A C, Ancona M, Ceolini E, et al. Towards Better Understanding of Gradient-based Attribution Methods for Deep Neural Networks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018:5-10

[16] Oquab M, Bottou L, Laptev I, et al. Is object localization for free-weakly-supervised learning with convolutional neural networks//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 685-694

[17] Selvaraju R R, Chattopadhyay P, Elhoseiny M, et al. Choose your neuron: Incorporating domain knowledge through neuron-importance//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 526-541

[18] Sharif Razavian A, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 806-813

[19] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 427-436

[20] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization//Proceedings of the 31st International Conference on Machine Learning. Lile, France,2015:1-12

[21] Bau D, Zhou B, Khosla A, et al. Network dissection: Quantifying interpretability of deep visual representations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Honolulu, USA,2017: 6541-6549

[22] David Bau, Jun-Yan Zhu, Hendrik Strobelt, et al. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks// Proceedings of the7th International Conference on Learning Representations, New Orleans, USA.2019:Poster.

[23] Cao C, Huang Y, Yang Y, et al. Feedback convolutional neural network for visual localization and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018, 41(7): 1627-1640.

[24] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 2668-2677

Ruan Li,Ph.D.,M.S. supervisor. Her main research interests include AI security, spatiotemporal data analysis, knowledge graph and distributed system

Wen Shasha, bachelor. Her main research interests focus on knowledge graph

Niu Yiming, undergraduate. His main research interests focus on knowledge graph.

Li Shaoning,undergraduate. His main research interests focus on timing analysis, network security.

XueYunzhi,Ph.D., professor. His research interests include trusted artificial intelligence, artificial intelligence testing and evaluation, and knowledge mapping.

RuanTao, master, assistantresearcher. His research interests include big data analysis, network security, domestic and foreign patent translation in the fields of computer, power, machinery, etc.

XiaoLimin, Ph.D., professor.His research interests include high performance computing and distributed system.

Background

The deep learning models have been widely used in computer vision, speech recognition and natural language processing for their advantages of deep layered learning and unlabeled learning models.

As a bottleneck restricting the further and long-term application of deep learning, the interpretability of deep learning models has been paid more and more attention. The deduction process of deep learning model learning and training is dominated by numerical operation and has a black box nature, which is lack of explainability and is difficult to understand through the concept of human society. Therefore, how to improve the transparency of deep neural network models has become a hot research topic.

The interpretability research of neural networks has experienced from the early coarse-grained exploration using the sensitivity difference characteristics of the model to the input, to the recent exploration of the function of single / combined neurons in a recognition task.

However, the existing research ideas are still only to analyze the decision process of neural network based on numerical results which lacks graphical representation of the overall understanding of neural network. Because the interpretation result of interpretable basis decomposition model for deep neural network is ofthe strict correspondence between scene and feature, it is a kind of semi-structured data, which has the advantage of easy to construct knowledge map. Therefore, this paper proposes a deep neural network visualization algorithm based on interpretable basis decomposition model and knowledge map. To the best of our knowledge, we are the first to combine interpretable basis decomposition model and knowledge map to deep network visualization to enhance its interpretability.

This work is supported by the National Key R&D Program of China under Grant NO.2017YFB0202004, the fund of the State Key Laboratory of Software Development Environment under Grant No.SKLSDE-2020ZX-15. National Natural Science Foundation of China (No. 11701545，No. 61772053).