卷积神经网络表征可视化研究综述

司念文[1] 张文林[1] 屈 丹[1] 罗向阳[2] 常禾雨[3] 牛 铜[1]

摘 要 近年来, 深度学习在图像分类、目标检测及场景识别等任务上取得了突破性进展, 这些任务多以卷积神经网络为基础搭建识别模型, 训练后的模型拥有优异的自动特征提取和预测性能, 能够为用户提供 "输入–输出"形式的端到端解决方案. 然而, 由于分布式的特征编码和越来越复杂的模型结构, 人们始终无法准确理解卷积神经网络模型内部知识表示, 以及促使其做出特定决策的潜在原因. 另一方面, 卷积神经网络模型在一些高风险领域的应用, 也要求对其决策原因进行充分了解, 方能获取用户信任. 因此, 卷积神经网络的可解释性问题逐渐受到关注. 研究人员针对性地提出了一系列用于理解和解释卷积神经网络的方法, 包括事后解释方法和构建自解释的模型等, 这些方法各有侧重和优势, 从多方面对卷积神经网络进行特征分析和决策解释. 表征可视化是其中一种重要的卷积神经网络可解释性方法, 能够对卷积神经网络所学特征及输入–输出之间的相关关系以视觉的方式呈现, 从而快速获取对卷积神经网络内部特征和决策的理解, 具有过程简单和效果直观的特点. 对近年来卷积神经网络表征可视化领域的相关文献进行了综合性回顾, 按照以下几个方面组织内容: 表征可视化研究的提起、相关概念及内容、可视化方法、可视化的效果评估及可视化的应用, 重点关注了表征可视化方法的分类及算法的具体过程. 最后是总结和对该领域仍存在的难点及未来研究趋势进行了展望.

关键词 深度学习, 卷积神经网络, 可解释性, 表征可视化, 显著图

引用格式 司念文, 张文林, 屈丹, 罗向阳, 常禾雨, 牛铜. 卷积神经网络表征可视化研究综述. 自动化学报, 2022, **48**(8): 1890−1920

DOI 10.16383/j.aas.c200554

Representation Visualization of Convolutional Neural Networks: A Survey

SI Nian-Wen[1] ZHANG Wen-Lin[1] QU Dan[1] LUO Xiang-Yang[2] CHANG He-Yu[3] NIU Tong[1]

Abstract In recent years, deep learning has made breakthrough progress on image classification, object detection, and scene recognition tasks. These tasks mostly build recognition models based on the convolutional neural network (CNN). The trained models have excellent automatic feature extraction and prediction performance, which is able to provide users with "input-output" end-to-end solutions. However, due to the distributed feature coding and the increasingly complex model structure, users cannot yet accurately understand the internal knowledge representation of the model as well as the potential reasons for a specific decision. On the other hand, the application of the CNN models in some high-risk areas also requires a full understanding of the reason for their decisions, so as to get user′s trust. Therefore, the interpreting ability of CNN has gradually attracted attention. Researchers have proposed a serious of methods for understanding and interpreting CNN, including post-hoc interpretation methods and building self-explainable models. These methods have their respective focuses and advantages, performing feature analysis and decision interpretation of CNN from various aspects. As one of the important CNN interpreting ability methods, representation visualization can visually present the features learned by CNN and the correlation between the input and output. In this way, a straightforward understanding of CNN internal features and decision-making can be obtained in a simple and intuitive way. This paper gives a comprehensive review of the related literatures on CNN representation visualization research in recent years, and organizes the content according to the following aspects: the introduction of representation visualization research, related concepts and contents, visualization methods, visualization effect evaluation, and the application of visualization. The classification of the representation visualization methods and the specific algorithms are our focus. Finally, the difficulties and future trends in the field are prospected, and the full text is summarized.

Key words Deep learning, convolutional neural networks, interpretability, representation visualization, saliency map

Citation Si Nian-Wen, Zhang Wen-Lin, Qu Dan, Luo Xiang-Yang, Chang He-Yu, Niu Tong. Representation visualization of convolutional neural networks: A survey. *Acta Automatica Sinica*, 2022, **48**(8): 1890−1920

收稿日期 2020-07-15 录用日期 2021-03-19
Manuscript received July 15, 2020; accepted March 19, 2021
国家自然科学基金 (61673395, U1804263) 和中原科技创新领军人才项目 (214200510019) 资助
Supported by National Natural Science Foundation of China (61673395, U1804263) and Zhongyuan Science and Technology Innovation Leading Talent Project (214200510019)
本文责任编委 王立威
Recommended by Associate Editor WANG Li-Wei

1. 信息工程大学信息系统工程学院 郑州 450001 2. 信息工程大学网络空间安全学院 郑州 450001 3. 信息工程大学密码工程学院 郑州 450001
1. College of Information System Engineering, Information Engineering University, Zhengzhou 450001 2. College of Cyberspace Security, Information Engineering University, Zhengzhou 450001 3. College of Cryptography Engineering, Information Engineering University, Zhengzhou 450001

近年来, 以深度神经网络 (Deep neural networks, DNN) 为代表的机器学习方法逐渐兴起[1]. 由于训练数据的增加[2−3] 及计算能力的大幅提升, DNN 的网络结构及与之相适应的优化算法[4−6] 变得更加复杂, DNN 在各项任务上的性能表现也越来越好, 产生了多种适用于不同类型数据处理任务的经典深度网络结构, 如卷积神经网络 (Convolutional neural network, CNN) 和循环神经网络 (Recurrent neural network, RNN). 对于图像数据处理与识别领域, CNN 是一种十分常用的网络结构, 在图像分类、目标检测、语义分割等任务上取得了非常好的效果, 已经成为该领域应用最广泛的基础模型[7].

如图 1 所示, 传统机器学习算法采用人工设计的特征集, 按照专家经验和领域知识将其组织到机器学习算法中. 由于设计人员本身了解这些被定义特征的具体含义, 因此, 传统机器学习方法一定程度上是可解释的, 人们大致明白算法对各种特征的依赖以及算法的决策依据. 例如, 线性模型可使用特征对应的权重代表特征重要程度. 相比于传统机器学习算法, 以 CNN 为代表的深度学习算法属于特征学习或表示学习, 可对输入数据进行自动特征提取及分布式表示, 解决了人工特征设计的难题. 这一优势使其能够学习到更加丰富完备的且含有大量深层语义信息的特征及特征组合, 因此在性能表现上超过多数传统机器学习算法.

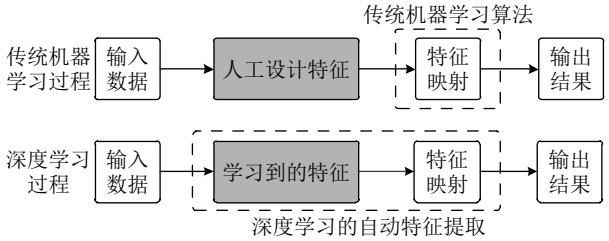


图 1　传统机器学习与深度学习的学习过程对比[8]

Fig. 1　Comparison of the learning process between traditional machine learning and deep learning[8]

然而, CNN 这一优势的背后也存在着一定局限性. 一方面, 人们至今无法较好地理解 CNN 内部知识表示及其准确的语义含义. 即使是模型设计者也难以回答 CNN 到底学习到了哪些特征、特征的具体组织形式以及不同特征的重要性度量等问题, 导致 CNN 模型的诊断与优化成为经验性、甚至盲目性的反复试探, 这不仅影响了模型性能, 还可能遗留潜在的漏洞; 另一方面, 基于 CNN 模型的现实应用在日常中已经大量部署, 如人脸识别、行人检测和场景分割等, 但对于一些风险承受能力较低的特殊行业, 如医疗、金融、交通、军事等领域, 可解释性和透明性问题成为其拓展和深入的重大阻碍. 这些领域对 CNN 等深度学习模型有着强烈的现实需求, 但受限于模型安全性与可解释性问题, 目前仍无法大规模使用. 模型在实际中可能犯一些常识性错误, 且无法提供错误原因, 导致人们难以信任其决策.

因此, 对 CNN 的理解和解释逐渐受到人们关注, 研究者们尝试从不同角度出发, 解释 CNN 的特征编码和决策机制. 表征可视化作为其中一种解释方法, 采用基于特征重要性的解释思路, 寻找输入变量、特征编码及输出结果之间的相关性, 并以视觉展示的方式直观呈现, 是一种较为直接的理解 CNN 的途径. 本文对该领域的现有研究进行了系统性整理和回顾, 对其中涉及的相关概念及内容、典型方法、效果评估、应用等方面作了归纳总结, 着重介绍了可视化方法的分类及算法的具体过程. 最后, 分析了该领域仍存在的难点并展望了未来研究趋势.

本文后续内容安排如下: 第 1 节简述了 CNN 表征可视化的相关概念和研究内容; 第 2 节梳理了现有的表征可视化方法, 对其进行了分类整理; 第 3 节介绍了部分可视化效果评估方法; 第 4 节简要阐述了可视化方法在一些领域的应用; 第 5 节探讨了该领域仍存在的难点及未来的研究趋势; 第 6 节总结全文.

1　相关概念与研究内容

1.1　相关概念

1.1.1　CNN

目前, CNN 已成为基于深度学习的图像识别领域应用最广泛、效果最佳的网络结构. 最早的 CNN 由 LeCun 等[9] 于 1998 年提出, 用于手写体数字识别. CNN 的基本结构中含有输入层、卷积层、全连接层及输出层. 其中输入层、全连接层、输出层与其他网络大致相同, 仅卷积层是 CNN 特有的结构. 经典 CNN 卷积层中含有卷积、激活和池化 3 种操作: 1) 卷积操作使用多个卷积核 (滤波器) 在输入张量上平移作内积运算, 得到对应的特征图. 同层的不同卷积核用来提取不同模式的特征, 不同层的卷积核则用来提取不同层级的特征. 2) 激活操作使用非线性激活函数处理卷积结果, 用于提升网络的非线性特性, 增强非线性拟合能力, 常用的激活函数如 tanh、sigmoid、rectified linear unit (ReLU)[6] 和改进版[10−11] 等. 3) 池化操作一般使用最大值池化

和平均值池化, 按照池化窗口处理整个窗口内的值, 用于压缩参数和降低过拟合.

稀疏连接和权重共享是 CNN 相对于前馈神经网络的主要特点. 基于这些经典的 CNN 结构及其特性, 研究人员通过不断改进和优化[12], 逐渐设计出结构更复杂且识别性能更优异的 CNN, 以在 Imagenet Large Scale Visual Recognition Competition (ILSVRC) 数据集[2] 图像分类任务上的优胜 CNN 模型为例:

1) 2012 年, Krizhevsky 等[1] 提出了 AlexNet, 在图像分类任务上以巨大优势取得冠军, 成功吸引了学术界的关注, 成为新阶段 CNN 兴起的标志.

2) 2013 年, Zeiler 等[13] 提出了 ZFNet, 利用反卷积可视化技术诊断 AlexNet 的内部表征, 然后对其针对性地做了改进, 使用较小的卷积核和步长, 从而提升了性能.

3) 2014 年, 谷歌公司 Szegedy 等[14] 提出了 GoogLeNet, 核心是其中的 Inception 模块, 使用了不同尺寸的卷积核进行多尺度的特征提取和融合, 从而更好地表征图像. 同年, 牛津大学的 Simonyan 等[15] 提出了视觉几何组网络 (Visual geometry group network, VGGNet), 仅使用 2×2 和 3×3 两种典型的卷积核, 通过简单地增加层的深度实现了性能提升.

4) 2015 年, 微软公司 He 等[16] 提出了残差网络 (Residual networks, ResNet), 使用残差连接实现跨层的信息传播, 缓解了之前由于深度增加引起的梯度消失问题, 并以 3.57% 的错误率首次超越人类水平.

5) 2016 年, Huang 等[17] 提出了 DenseNet, 相比于 ResNet, 使用了密集连接操作, 强化特征的传播和复用.

6) 2017 年, Hu 等[18] 提出了压缩激励网络 (Squeeze-and-excitation networks, SENet), 通过特征图各通道间的权值自适应再调整, 实现各个通道之间的特征重标定, 提升了网络的特征提取能力.

CNN 在图像数据处理上有天然的优势, 因而在图像分类、目标检测、语义分割和场景识别等领域应用广泛, 在其他模态的数据如视频、语音和文本等领域也有较多应用. 图像分类是 CNN 最典型的应用领域, 许多图像分类系统使用预训练的 CNN 进行部署. 预训练的 CNN 是指已经在某个数据集上完成训练的 CNN 模型. 一般情况下, 预训练的 CNN 由研究人员设计并调整至最佳状态, 在实际场景中可以直接使用而无需再训练. 由于预训练 CNN 模型在现实中经常使用, 因此, 针对预训练 CNN 模型的理解和解释是可解释性研究中的一项重要内容.

1.1.2 可解释性

可解释性是近年来深度学习领域的研究热点. 可解释性与可理解性的含义并不相同[19−20], 文献 [19] 从 CNN 特征表示形式的角度出发, 对 CNN 的 “可解释性”和 “可理解性”做了区分: 可解释性表示从抽象概念 (向量空间、非结构化特征空间) 到人类可理解的领域(图像和文字等) 的映射, 而可理解性表示可解释域内促使模型产生特定决策的一组特征. 从这种区分看, “可解释性”研究重点在于将参数化形式表示的特征映射到人类可直观感受的表示形式, 而 “可理解性”侧重在人类可理解的领域中寻找与模型某个决策相关的具体特征. 也就是说, “解释”是一种从不可解释域到可解释域的映射动作, “理解”则是一种在可解释域内寻找感兴趣证据的过程. 麻省理工的研究人员认为[20], 通过 “解释”能够实现对深度网络的 “理解”, 可解释性的研究目标是以某种人类可理解的方式描述一个系统的内部机制. 同时, 将可解释性的研究内容分为 DNN 处理过程的理解、DNN 内部表征的理解和自解释的 DNN 三个方面.

深度学习可解释性的研究内容非常丰富, 本文从可解释性研究的模型对象出发, 根据待解释的目标模型是否已经完成训练, 将深度学习可解释性研究划分为两部分: 事后解释和自解释模型, 如图 2 所示[21].

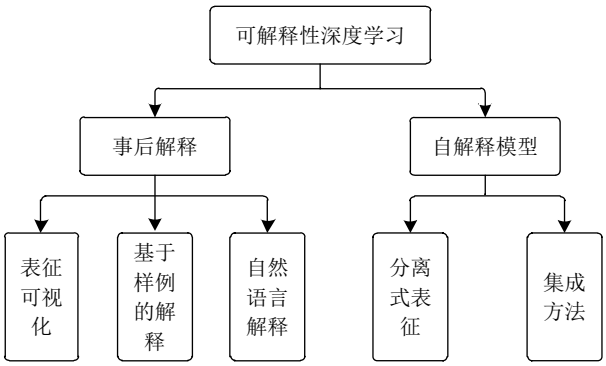


图 2 可解释性深度学习的研究内容划分

Fig. 2 The division of the research content of the interpretable deep learning

事后解释是对预训练模型的解释. 现实中, 由于模型已经完成训练和部署, 而重新训练模型耗费的时间和资源成本过大, 因此不具备重新训练的可能性. 针对这种模型的解释, 需要在不修改模型自身结构及参数的情况下完成, 结合预训练模型的输入、中间层参数和输出等信息, 实现对模型内部表征及决策结果的解释.

对于预训练模型的事后解释方法, 现有研究主要分为以下 3 类:

1) 表征可视化. 表征可视化是一种基于特征重要性的解释方法, 主要研究模型内部的特征表示及这些特征与模型输入、输出之间的关系. 梯度归因方法[22−23] 是最具代表性的表征可视化方法, 使用输入空间中像素自身的梯度 (或绝对值、平方等) 来衡量该像素与预测结果的关联程度. 表征可视化与模型结构可视化不同, 前者重在研究模型内部特征 (以参数的形式) 的语义理解, 以及输入、特征编码及输出之间的因果关系, 后者研究模型结构、数据流向及形状的变化.

2) 基于样例的解释. 基于样例的解释是一种基于样本重要性的解释方法, 采用训练数据中的样本原型作为当前决策结果的解释[24−25]. 这种方法模拟人对事物的解释过程[26], 从数据集中已有样本 (已经学习过) 中找到相似样本, 作为对新的样本预测结果的比较.

3) 自然语言解释. 自然语言解释以人类可理解的自然语言形式, 对 CNN 识别结果进行解释[27]. 该过程中, 需要将 CNN 的图像特征编码映射为 RNN 的自然语言特征编码, 通过跨模态的表征融合来生成用于解释 CNN 输入与输出的自然语言. 该过程与图像描述[28] 和视觉问答[29] 相似.

自解释模型不同于事后解释, 其在模型设计时即考虑了内在可解释性, 在此基础上进行训练和优化, 形成结构上或逻辑上具有内生可解释性的模型. 自解释模型能够在应用的同时由其自身为用户提供对输出结果的解释.

对于建立具有自身可解释性的模型, 现有研究主要分为以下 2 类:

1) 分离式表征: 在模型结构或优化过程中添加一些约束, 以降低模型复杂性, 同时保证模型的性能, 使模型内部的表征分离可理解. 例如, Zhang 等[30] 对滤波器的学习进行约束, 训练出可解释的滤波器, 使每个滤波器有针对性地关注特定目标部位.

2) 集成方法: 结合传统可解释性较好的机器学习方法, 构建在深度神经网络的识别性能和传统方法的可解释性之间折衷的新模型. 例如, 将神经网络集成到决策树算法中, 使用神经网络提取的特征作为输入, 这样训练得到的模型同时具有两者的优点, 可实现决策路径的清晰可理解[31].

1.1.3　表征可视化

表征可视化是一种事后解释方法, 通常以视觉的方式对 CNN 内部表征和输出决策进行解释. 表征可视化尝试解释 CNN 内部特征的表示形式、输入–内部特征–输出三者之间的关系、促使网络做出当前预测的输入等问题。与其他方法相比, 表征可视化方法具有以下优点: 1) 简单直观, 从视觉上为用户提供观察. 2) 便于深度分析网络表征, 诊断训练效果, 进而改进网络结构设计. 3) 无需修改模型结构, 多数表征可视化方法可在模型完成训练之后进行特征分析与决策结果解释, 无需修改或重新训练模型. 表征可视化方法生成的解释结果以热力图的方式呈现. 热力图是一个由不同颜色强度构成的图像, 像素颜色的强度与其重要性相对应. 从数学角度看, 热力图实际上是一组与输入变量对应的重要性值 (或相关性值) 的集合, 集合中的每个元素值表示其对应的输入变量与输出结果之间的相关性.

1) CNN 表征可视化

表征可视化过程与 CNN 预测过程相互依赖, 如图 3 所示. 图 3 上方为 CNN 预测过程, 下方为可视化方法的解释过程, 箭头表示这两个过程中各阶段之间的相互关系.

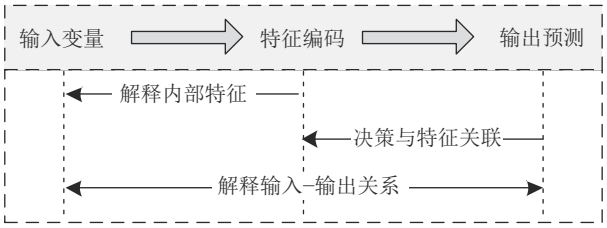


图 3　CNN 表征可视化的研究思路

Fig. 3　The research idea of CNN representation visualization

CNN 预测过程: 实现从输入变量到输出预测的映射. 其中, 输入变量对应的输入空间被认为是人类可理解的空间 (例如图像和语言文本), 而特征编码对应的特征空间经过了 CNN 的自动特征提取与特征组合. 可视化解释 CNN 的目的就是将中间层特征编码和输出层预测结果反向映射到输入空间, 实现不可解释域向可解释域的映射.

可视化方法的解释过程涉及 3 种: 1) 解释内部特征: 研究黑盒中间编码了哪些知识, 以怎样的形式组织这些知识的. 2) 决策与特征关联: 研究中间层的知识与输出预测之间的关系. 3) 解释输入–输出关系: 研究输入变量、中间层特征编码和输出预测三者之间的关系.

2) CNN、RNN 和生成对抗网络表征可视化的比较

CNN 在图像数据处理领域应用较为广泛, 层次化的表征方式使其适用于图像数据逐层学习的特性, 与人类非常相似. 因此, CNN 表征可视化主要

研究各个隐含层所编码的特征、这些特征的语义含义及与输入输出之间的关系. 对于另外两种常见的DNN: 循环神经网络 (RNN) 与生成对抗网络 (Generative adversarial network, GAN), 表征可视化研究的关注点略有不同.

RNN 是一种随时间步迭代的深度网络, 有长短时记忆网络 、门控循环单元等扩展版结构, 擅长处理时序型数据, 在自然语言处理领域应用广泛. RNN 的主要特点在于其迭代式的处理数据, 这些迭代信息存储于网络结构中的隐状态中, 每个时间步的隐状态含义不同, RNN 的长距离依赖关系学习能力也在于这些隐状态的学习效果. 因此, RNN 可视化研究多专注于对这些隐藏状态的理解与解释. 例如, 文献 [32] 可视化 RNN 的隐状态对于输入的预期响应, 用于观察 RNN 内部的正面与负面输入时的激活分布. 文献 [33] 开发了一个长短时记忆网络可视化工具, 用于了解这些隐藏状态的动力学过程. 文献 [34] 通过可视化的方式解释了长短时记忆网络在长距离依赖关系学习上的优势. 此外, 一些图像领域常用的表征可视化方法如层级相关性反馈 (Layer-wise relevance propagation, LRP) 方法, 也被用于解释 RNN 的表征及量化输入–输出之间的关系[35−36].

GAN 是一种生成式神经网络, 由生成器和判别器两部分构成, 二者之间通过对抗学习的方式互相提升性能[37]. 从结构上看, GAN 的生成器一般使用反卷积结构, 判别器可视为一个 CNN 结构. 由于 GAN 主要用于学习数据的潜在分布, 然后用于生成式任务, 因此, GAN 可视化的关注点主要在于生成器部分. 更具体地, 在于理解和解释生成器隐变量的作用. 典型的如 InfoGAN[38], 对输入向量进行分解, 使其转为可解释的隐变量及不可压缩的噪声, 进而约束隐变量与输出之间的关系, 从而学习可解释的特征表达. 文献 [39] 和文献 [40] 通过操纵生成器的隐变量来观察生成结果的变化情况, 进而理解 GAN 的过程. 文献 [41] 专门研究了 GAN 隐空间的语义解纠缠问题, 提出了一种效果较好的人脸编辑方法, 可通过编辑 GAN 的隐空间来调整生成人脸的属性, 如姿势、性别和年龄等.

1.2 研究内容

本文梳理了 CNN 表征可视化的研究内容, 如图 4 所示, 主要分为以下 3 个方面:

1) 可视化方法. 从不同的目标模型和解释需求出发, 研究不同侧重点的表征可视化方法, 从而提升可视化解释的效果.

2) 可视化效果的评估. 研究可视化效果的评估方法, 主要从两个方面展开: 有效性评估和鲁棒性评估. 有效性评估用于评价可视化方法的解释效果, 分别从定性和定量的角度进行度量. 鲁棒性评估用于评价可视化方法在对抗性输入的作用下能否有效地提供合理的解释.

3) 可视化的应用. 研究可视化方法在相关领域的应用, 根据不同可视化方法的特点为其选择合适的应用场景, 例如用于诊断网络缺陷、为模型提供面向重要特征的注意力机制、用于弱监督目标定位任务等.

2 可视化方法

2.1 方法分类

根据可视化方法的算法原理, 可归纳为 6 种主要类型: 基于扰动的方法、基于反向传播的方法、类激活映射、激活最大化、注意力掩码和其他方法. 下面分别对每类方法进行介绍.

2.1.1 基于扰动的方法

基于扰动的可视化类似于一种因果过程, 通过修改输入 (原因) 观察输出 (结果) 的变化情况, 从而确定被修改的输入对于输出的影响大小. 如图 5 所示, 考虑输入图像 $\boldsymbol{x} \in \mathbf{R}^d$, 分类器 f, 输出结果为 $f(\boldsymbol{x}) \in \mathbf{R}^n$, 其中类别 c 的 Softmax 分数为 $f^c(\boldsymbol{x})$. 扰动方法研究使用删除、遮挡、模糊等方式处理 $\boldsymbol{x}$ 的最小区域, 观察 $f^c(\boldsymbol{x})$ 的变化. 若 $f^c(\boldsymbol{x})$ 下降较大, 则被遮挡区域对 $f^c(\boldsymbol{x})$ 影响较大, 重要性也较大.

典型的扰动方式分为以下 3 种:

1) 简单扰动. 文献 [13] 使用固定尺寸的像素块 (如 2 × 2), 按照从左到向、从上到下的顺序依次遮挡图像的各区域, 观察各遮挡后图像的预测结果. 针对特定预测类别, 分数下降越大, 表明此时被遮挡区域对于该类别越重要, 从而形成基于像素块重要性的显著图. 文献 [42] 认为使用随机值作为遮挡模块的填充像素更合理, 而不应仅使用灰色像素. 文献 [43] 使用蒙特卡洛抽样产生多个扰动掩码 $\boldsymbol{M}_i$ 用于遮挡图像 (见图 6), 利用 f 对扰动图像 $\boldsymbol{x} \odot \boldsymbol{M}_i$ 进行分类, 从而得到置信度向量, 表示该掩码与各个类别的关联度大小. 最后, 使用类别对应的置信度对各掩码作线性加权, 得到最终的掩码.

2) 有意义的扰动. 文献 [44] 认为用于遮挡图像的掩码是可以学习的, 而不需要使用简单的平移或随机等方式盲目的遮挡. 使用优化思想学习到的像素级扰动掩码有明确的含义, 可以更有效地遮挡重

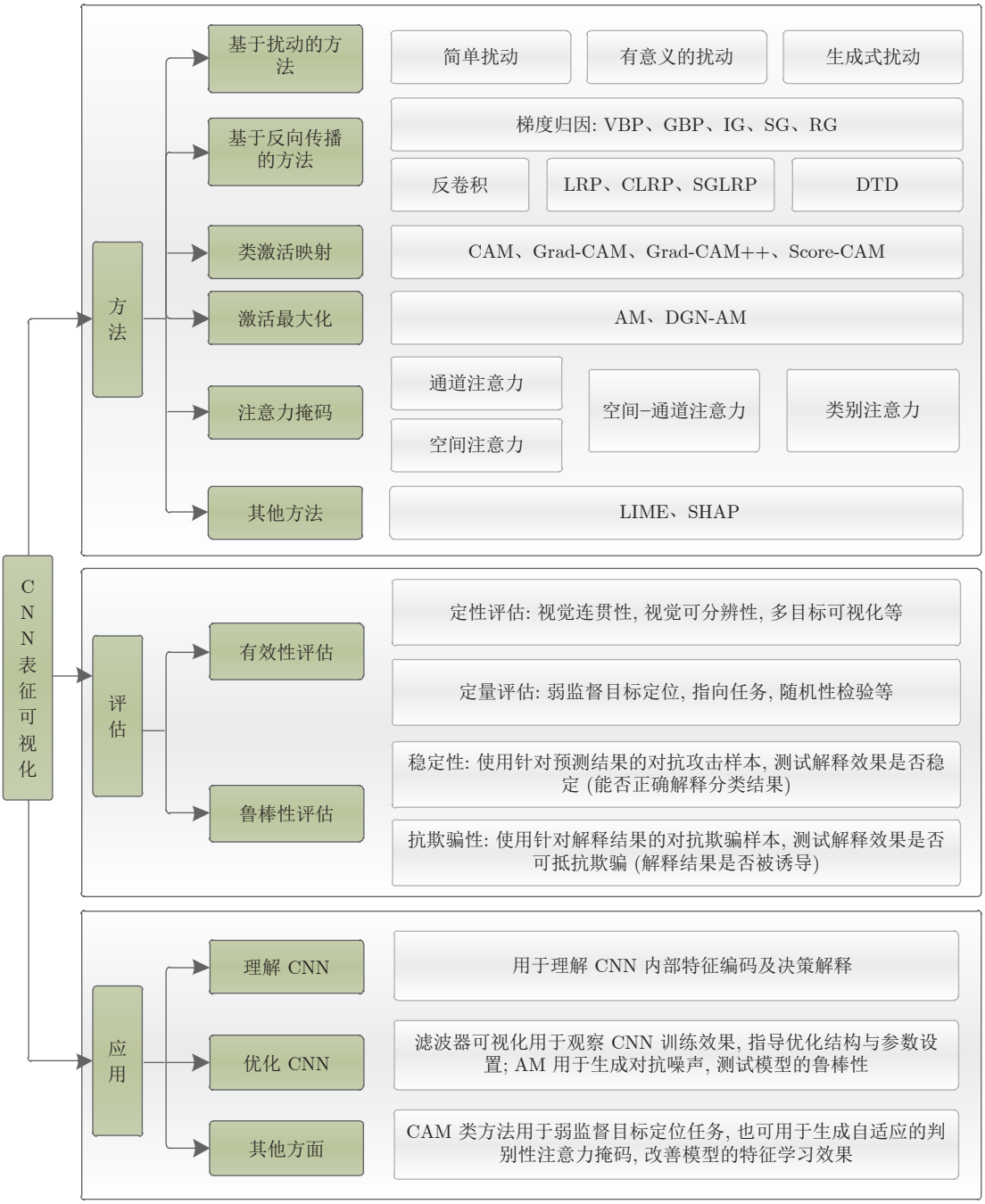


图 4　CNN 表征可视化的研究内容

Fig. 4　Research content of the CNN representation visualization

要像素, 使用的扰动方式如下[44]:

$$[\Phi(x_0;m)](u)=\begin{cases}m(u)x_0(u)+(1-m(u))u_0, & \text{常数}\\ m(u)x_0(u)+(1-m(u))\eta(u), & \text{噪声}\\ \int g_{\sigma_0 m(u)}(v-u)x_0(v)\mathrm{d}v, & \text{模糊}\end{cases}\tag{1}$$

式中, $x_0(u)$ 表示输入图像 (自变量为元素位置 u), $m(u)$ 表示掩码. 第 1 种使用常数扰动, u_0 表示色彩均值; 第 2 种使用噪声扰动, $\eta(u)$ 表示高斯噪声; 第 3 种使用模糊扰动, σ_0 表示高斯模糊核 g_σ 的标准差. 定义目标函数如下:

$$m^*=\arg\min_m \lambda||1-m||_1+f^c(\Phi(x_0;m))\tag{2}$$

使用优化方式学习到的扰动掩码, 对于指定的目标类别, 可以使掩码遮挡后的图像的预测分类达到局部最低, 即表示掩码有效遮挡了图像中最重要

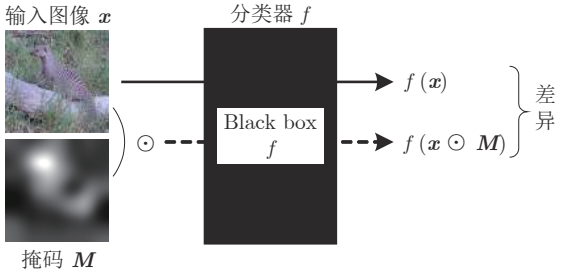


图 5 基于扰动的方法的解释流程

Fig. 5 Interpretation process of the perturbation based method

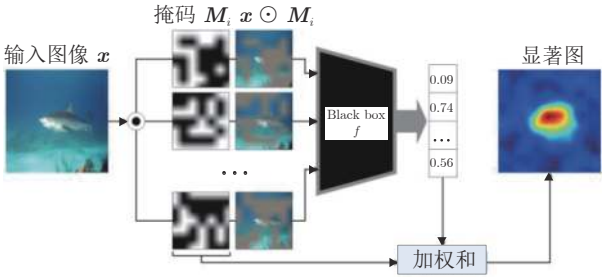


图 6 使用随机采样产生扰动掩码的过程[43]

Fig. 6 The process of generating a perturbation mask using random sampling[43]

的区域. 该式同时使用 L1 范数约束掩码, 使扰动区域尽可能小, 以实现扰动最少最关键像素的目的.

3) 生成式扰动. 文献 [45] 和文献 [46] 使用生成模型扰动输入特征, 得到视觉上更加自然的扰动图像. 如图 7(d) 所示, 与图 7(b) 模糊、图 7(c) 灰度化仍保留着目标轮廓相比, 生成式模型修复的图 7(d) 在目标位置插入与环境一致的平滑像素特征, 可使扰动图像在视觉上仍然是一张自然图像, 而非经过了明显地遮挡.

此外, 还有极值扰动[47] 和对抗性扰动[48] 等方法. 总体来看, 简单扰动方法将目标网络当作黑盒, 仅需获取其 "输入–输出"对, 无需接触网络权重与中间层激活, 解释过程所需的资源较少, 但其解释

图 7 使用生成式模型生成扰动[45] ((a) 原图, (b) 模糊; (c) 灰度; (d) 生成扰动; (e) 随机噪声)

Fig. 7 Using generative models to generate perturbation[45] ((a) Original image; (b) Blur; (c) Gray; (d) Generated perturbation; (e) Random noise)

效果欠佳. 更复杂的扰动方法则需要多次优化迭代, 所需时间较长.

2.1.2 基于反向传播的方法

基于反向传播的方法根据所设计的反向传播规则, 将网络输出层预测值逐层分解并传播到输入空间, 从而确定输入–输出之间的相关性. 输出层信息回传过程中, 利用模型权重参数作为引导, 结合网络中神经元的正向激活值, 进行逐层反向计算, 直到输入空间, 为每个变量 (或一组变量) 分配与输出预测相关的贡献值.

图 8 描述了基于反向传播的可视化方法的解释流程. 对于输入图像 $\boldsymbol{x}=\{x_1,\cdots,x_d\}$, d 表示输入图像对应矢量的维度, x_i 表示该矢量的第$i(1\leq i\leq d)$维. 设目标网络为 f, 解释方法为 g, 解释过程可形式化为:

$$R(\boldsymbol{x})=\{R(x_1),\cdots,R(x_d)\}=g(\boldsymbol{x};f) \tag{3}$$

式中, $R(x_i)$ 表示输入的第 i 维元素与输出预测 $f(\boldsymbol{x})$ 之间的相关性值.

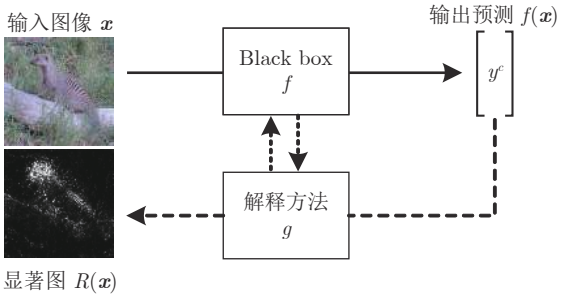


图 8 基于反向传播的方法的解释流程

Fig. 8 Interpretation process of the backpropagation based method

2.1.2.1 梯度方法及其变种

1) 普通反向传播 (Vanilla backpropagation, VBP). 文献 [22] 和文献 [23] 提出最简单的基于梯度的可视化方法, 使用网络输出得分的输入空间中数据点的导数, 表示该点与输出结果之间的相关性大小, 形式化如下:

$$\boldsymbol{M}_{sm}^{c}(x_i)=\frac{\partial f^c(\boldsymbol{x})}{\partial x_i} \tag{4}$$

式中, $\boldsymbol{x}$ 表示输入图像, $f^c(\boldsymbol{x})$ 表示对 $\boldsymbol{x}$ 的预测类别 c 的得分. 如图 9 所示, 表示在 AlexNet[1] 上使用 VBP 方法进行可视化的过程. 其中, C 表示卷积层, FC 表示全连接层.

VBP 方法的反向传播过程基于链式法则进行, 由输入图像的梯度构成显著图, 显著图中较亮位置的梯度绝对值也越大, 这些突出位置显示了与模型

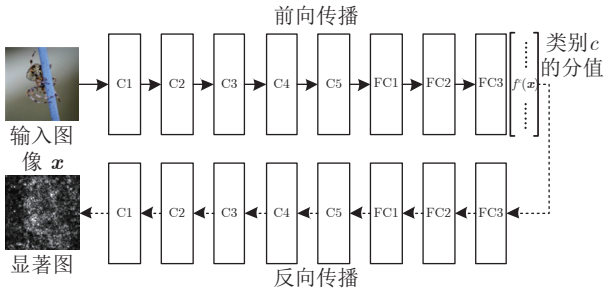


图 9　VBP 方法的过程[49]

Fig. 9　The process of the VBP method[49]

输出结果相关的输入特征, 且相关性越高则显著图中对应特征越明显. 直观理解, 梯度表示模型训练中参数的更新量, 数据点的梯度越大表示输出对该点的变化越敏感, 则该点与输出的相关性也越大. 因此, 这种方法也称作灵敏度分析方法, 用于分析输出结果对输入特征的变化量的敏感程度. 初始得到的显著图, 通过取绝对值或取平方等后处理策略, 使热力图中显著性区域更加集中, 散点噪声更少.

2) 导向反向传播 (Guided backpropagation, GBP). VBP 方法基于普通的梯度反向传播, 依靠网络自身反向传播所产生的梯度, 仅对最终的梯度图进行后处理, 而未对反向传播过程作任何更改. GBP 方法[50] 与其略有不同, 它对反向传播中每个卷积层的梯度都进行负值过滤, 使负值调整为 0, 即使网络中间有 ReLU 层, 也会对其梯度再进行一次 ReLU 过滤. 这样做的目的是仅保留与网络输出正相关的梯度, 去掉负相关的梯度.

VBP 和 GBP 是 2 种最简单的梯度方法, 但由于深层网络的梯度传播过程中的问题, 导致这两种方法也含有一定的局限性, 体现在以下 3 个方面:

1) 梯度消失. 随着反向传播过程层数的增加, 梯度分布越来越稀疏, 再加上一些梯度过滤措施, 导致显著图愈发不明显. 该问题在很多深层网络中尤其突出.

2) 梯度不能完全反映输入特征的重要性. 梯度用于表示对应点的变化量对输出的影响, 但不一定可以解释该点自身对输出的贡献. 例如, 沿着梯度下降方向改变 (增加/减小) 该点的数值能够使输出结果分值更高, 但并不表示该点自身对输出有较大影响. 因此, 梯度方法并不能完全理解为能够解释 CNN 的分类结果, 仅能解释怎样改变输入可使该分类结果的置信度更高.

如图 10 所示, 左侧表示输入图像, 右侧表示显著图的解释结果. 在第 1 行中, 梯度突出的区域偏向左上方, 表明该区域对 “reflex camera”标签更重要, 而原图中对应区域并没有实际含义. 在第 2 行中,去除这部分区域后, 在预测类别不变且置信度增加的情形下, 梯度突出的区域却发生了改变. 从中可以看出, 对于一些特殊的图像, 梯度自身能否表明像素重要性值得怀疑.

图 10　梯度不稳定导致解释结果的不确定性[51]

Fig. 10　Uncertainty of interpretation results due to gradient instability[51]

3) 梯度噪声. 梯度对应的显著图中, 散点噪声较多, 只能从整体上观察显著图中较模糊的目标信息. 目前的研究对这些噪声的出现作了一些假设, 但仍有待进一步探讨. 图 11 表明了梯度显著图的噪声问题, 与给定的边框相比, 显著图中高亮区域含有较多的框外散点噪声.

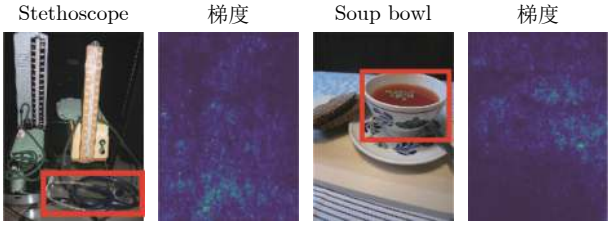


图 11　梯度方法产生的显著图含有大量噪声[44]

Fig. 11　The saliency map generated by the gradient method contains a lot of noise[44]

尽管如此, 基于梯度的显著图可视化仍有一些有益的作用, 是一种能够有效辅助理解 CNN 的方法. 针对上述问题, 一些研究尝试使用新的方法对梯度反向传播规则或梯度图自身进行处理, 以提升梯度显著图的可视化效果, 形成像素区域更加集中的和连续的梯度图. 这些研究包含以下几项:

1) 平滑梯度. Smilkov 等[52] 对梯度图中噪声出现的原因作了假设, 认为这些噪声是由神经网络学习到的得分函数的不平滑性引起的, 噪声即是一些变量的梯度对应的没有实际含义的局部变化. ReLU 激活函数就是一种典型的不平滑激活函数. 图 12 表示在不改变图像的视觉效果与分类结果的情况

下, 对单个输入像素添加细微扰动, 扰动率 (横轴) 的增加引起的该像素各通道的梯度值 (纵轴) 的变化. 可以看出, 像素梯度变化对像素的改变具有较大敏感性, 尽管这种改变并不一定具有实际含义.

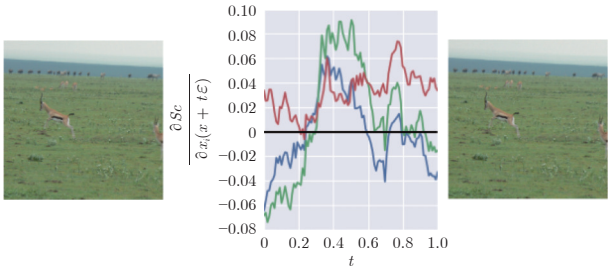


图 12 单个像素的梯度值的不稳定性[52]

Fig. 12 The instability of the gradient value of a single pixel[52]

梯度的不稳定特性是 “平滑梯度”方法提出的依据, 通过为输入图像添加采样自高斯分布的噪声, 对每次生成的梯度图叠加后再平均, 实现对梯度图的平滑与去噪. 具体地, 对于输入图像 x, 添加服从高斯分布 $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ 的噪声, 形成在视觉上无明显变化的 N 张带噪声图像 $x+\varepsilon$, 然后对这些图像生成的梯度图 $\boldsymbol{M}^c_{sm}(x+\varepsilon)$ 加以平均, 得到最终的显著图 $\boldsymbol{M}^c_{sg}(x) = \sum_{n=1}^{N} \boldsymbol{M}^c_{sm}(x+\varepsilon)$. 这样通过多个显著图平均化的方式实现图像的平滑, 从而去除显著图中存在的噪声, 得到更好的可视化效果.

2) 积分梯度[53]. 将输入图像从某个初始值 (例如 0) 开始, 按照一定比例放大到当前值. 将中间过程中的每个输入对应的显著图进行平均, 得到最终的显著图.

3) 整流梯度. Kim 等[54] 深入对比了现有的梯度反向传播方法, 并提出假设认为, 由于 CNN 在前向传播中本身含有噪声, 导致反向传播中相应元素位置会出现梯度噪声. 为了处理掉显著图中的噪声, 通过设置适当的阈值, 使重要性分数超过该阈值的神经元的梯度才会被反向传播, 从而过滤非重要的特征.

此外, 文献 [55] 还提供了梯度归因方法的比较基准. 文献 [56] 对几种基于梯度的可视化方法的效果作了详细对比, 引入了一种量化评价标准 *sensitivity-n* 统一进行度量. 文献 [57] 采用聚合的思路, 对几种方法的结果进行聚合, 实现更加稳定的解释. 总体来看, 梯度方法及其变种主要依靠基于梯度的反向传播实现解释过程, 各种方法在传播规则上略有不同, 表 1 对比了各方法的特点.

表 1 梯度方法及其变种的特点比较

Table 1 Comparison of the characteristics of the gradient method and its variants

方法	显著图生成依据	特点
VBP	普通梯度	过程简单, 但存在梯度噪声问题
GBP	每一层使用 ReLU	过程简单, 但存在梯度噪声问题
积分梯度	梯度图的平均	过程复杂, 需多次迭代, 耗时
平滑梯度	梯度图的平均	过程复杂, 需多次迭代, 耗时
整流梯度	阈值过滤后的梯度	过程较复杂, 阈值的选取需要经验

2.1.2.2 反卷积

Zeiler 等[13] 提出基于反卷积的可视化方法, 反向重建 CNN 中间层神经元学习到的模式. 通过正向卷积的逆过程, 将中间层激活值逐层反向卷积到输入空间, 在输入空间找到激活该神经元的特征.

为了清楚介绍经典反卷积过程, 首先将正向卷积过程形式化. 如图 13 所示, 对于正向卷积[1, 15], 输入图像 $\boldsymbol{x}$ 经过含有 L 个卷积层的网络 f, 进行特征提取与分类, 该过程形式化表示如下:

$$(\boldsymbol{A}^1, \boldsymbol{A}^2, \cdots, \boldsymbol{A}^L) = f(\boldsymbol{x}; \theta) \tag{5}$$

式中, $\boldsymbol{A}^l$ $(1 \le l \le L)$ 表示中间第 l 个卷积层输出的多通道特征图, θ 表示 CNN 的参数集合, 函数 f 表示正向卷积过程.

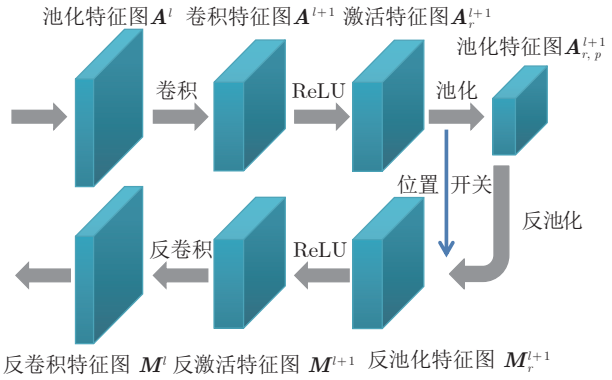


图 13 反卷积可视化方法的过程

Fig. 13 The process of deconvolution visualization method

1) 卷积层: 将卷积过程形式化为第 l 层的卷积核 $\boldsymbol{f}^l$ 作用于一个多通道特征图 $\boldsymbol{A}^l$, $*$ 表示卷积运算, 如下:

$$\boldsymbol{A}^{l+1} = \boldsymbol{A}^l * f^l \tag{6}$$

2) 激活层: 使用激活函数 ReLU 过滤负激活值, 该过程如下:

$$\boldsymbol{A}^{l+1}_r = \mathrm{ReLU}(\boldsymbol{A}^{l+1}) \tag{7}$$

3) 池化层: 使用最大值池化过程中, 仅取每个池化窗口中的最大值, 并纪录各最大值在整张特征图中的位置, 称为 switches. 它可以在反向池化过

程中, 按照纪录的位置恢复最大值. 过程如下:

$$\boldsymbol{A}_{r,p}^{l+1} = \text{Maxpool}(\boldsymbol{A}_r^{l+1}) \tag{8}$$

对于反卷积过程, 形式化如下:

1) 反池化层: 按照正向最大值池化中纪录的最大值位置, 将反向过程中的特征图 $\boldsymbol{M}_r^l$ 各元素恢复到其原始位置. 该过程增加了特征图尺寸, 恢复到正向池化前的特征图尺寸. 对于在正向池化中被丢弃的值所在的位置, 在反向池化中对其填充 0. 过程如下:

$$\boldsymbol{M}_r^{l+1} = \text{UnMaxpool}(\boldsymbol{M}_{r,p}^{l+1}) \tag{9}$$

2) 反激活层: 使用激活函数 ReLU 处理反池化得到的特征图:

$$\boldsymbol{M}^{l+1} = \text{ReLU}(\boldsymbol{M}_r^{l+1}) \tag{10}$$

3) 反卷积层: 使用正向卷积核的转置作为该层的反向卷积核:

$$\boldsymbol{M}^l = \boldsymbol{M}^{l+1} * (f^l)^T \tag{11}$$

上述过程表明, 反卷积和正向卷积相似, 可对特征图连续处理, 从中间层恢复到输入层, 实现上采样效果. 该过程中, 仅逐层恢复较强的激活值, 一直到输入空间, 这样可以发现输入空间中哪些特征引起了中间层某些较强的激活.

VBP、GBP 和反卷积 3 种典型的反向传播方法过程大致相同, 但在反向传播中对 ReLU 函数的处理策略却不同. 图 14 是 3 种方法的反向传播过程对比, 使用了 ReLU 或 ReLU$^{\text{BP}}$ 函数. 其中, Conv 表示正向过程卷积, Conv$^{\text{BP}}$ 表示对应的反向过程. ReLU$^{\text{BP}}$ 表示激活值为正, ReLU$^{\text{BP}}$ 表示激活值为正的区域反向传播其梯度, 激活值为负或零的区域其反向梯度为零. ReLU 则根据梯度值自身的正负来过滤梯度. GBP 综合了 ReLU和 ReLU$^{\text{BP}}$, 只有在激活值和梯度值均为正的情形下, 该位置的梯度才会反向传播, 其余位置仍保持为零.

2.1.2.3　LRP 及其变种

Bach 等[58] 提出层级相关性反馈 (Layer-wise relevance propagation, LRP), 重新定义了新的反向传播规则, 使用相关性函数 R 来计算某个输入变量对函数值的贡献, 如图 15 所示. 其中, 函数值 $f(\boldsymbol{x})$ 可按照制定的逐层相关性传播规则, 从输出层一直分解到输入空间的每个变量上, 以度量每个变量与函数值之间的相关性. $R(\boldsymbol{x})$ 表示输入变量 $\boldsymbol{x}$ 的每一维对该输出的贡献大小.

如图 16 所示, 假设输入层向量记作 $\boldsymbol{x} = (x_d)_{d=1}^V$, 其中 $1 \le d \le V$, V 表示输入层向量维度. 中间第 l 层向量记作 $\boldsymbol{z} = (z_d^{(l)})_{d=1}^{V(l)}$, $V(l)$ 表示第 l 层向量维

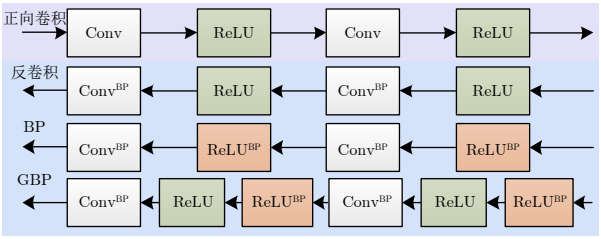


图 14　VBP、GBP 和反卷积三者之间的关系[49]

Fig. 14　The relationship of VBP, GBP and deconvolution[49]

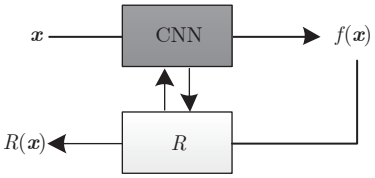


图 15　LRP 的过程

Fig. 15　The process of the LRP

度, 对应的激活值记作 $a = (a_d^{(l)})_{d=1}^{V(l)}$, 如下[19]:

$$a_k = \sigma(z_k) = \sigma\left(\sum_j a_j w_{jk} + b_k\right) \tag{12}$$

将输入层记作第 0 层, 中间层从第 1 层开始, 最后一层为第 L 层.

在反向传播中, 各层神经元的相关性分值定义如下: 将输入层变量的相关性分值记作 $R(\boldsymbol{x}) = R(x_d)_{d=1}^V$, 简写为 $(R_d)_{d=1}^V$. 中间第 l 层神经元的相关性分值记作 $R(\boldsymbol{z}) = R(z_d^{(l)})_{d=1}^{V(l)}$, 简写为 $R(\boldsymbol{z}) = (R_d^{(l)})_{d=1}^{V(l)}$.

LRP 方法中涉及几个约束条件如下[19]:

约束 1. 输入空间中变量贡献值的正负由其相关性分值的正负决定. 相关性函数值 $R(x_i)$ 的含义如下:

$$\begin{cases} x_i\text{对}f(\boldsymbol{x})\text{的贡献为正}, & R(x_i) > 0 \\ x_i\text{对}f(\boldsymbol{x})\text{的贡献为负}, & R(x_i) < 0 \end{cases} \tag{13}$$

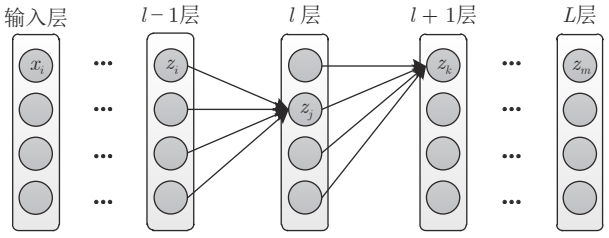


图 16　LRP 正向传播的过程[19]

Fig. 16　The forward propagation process of the LRP[19]

约束 2. 输入空间贡献值由函数值 $f(\boldsymbol{x})$ 分配. $f(\boldsymbol{x})$ 为网络输出层目标神经元的函数值, 其值等于输入空间所有变量的相关性值之和:

$$f(\boldsymbol{x}) = \sum_{d=1}^{V} R_d \tag{14}$$

约束 3. 层级相关性分值守恒. 函数值 $f(x)$ 对中间各层神经元的相关性值分配满足守恒定律:

$$f(\boldsymbol{x}) = \cdots = \sum_{d=1}^{V(l+1)} R_d^{(l+1)} = \sum_{d=1}^{V(l)} R_d^{(l)} = \sum_{d=1}^{V(l-1)} R_d^{(l-1)} = \cdots = \sum_{d=1}^{V(1)} R_d^{(1)} \tag{15}$$

约束 4. 第 l 层某个神经元被分解出的相关性值, 等于其流向的第 $l-1$ 层中所有神经元的相关性值之和:

$$R_j^{(l)} = \sum_{i=1}^{V(l-1)} R_{i\leftarrow j}^{(l-1,l)} \tag{16}$$

约束 5. 第 l 层某个神经元被流入的相关性值, 等于第 $l+1$ 层所有流向该神经元的相关性值之和:

$$R_j^{(l)} = \sum_{k=1}^{V(l+1)} R_{j\leftarrow k}^{(l,l+1)} \tag{17}$$

上述定义和约束规定了相关性值的含义及守恒原则, 对应的示例如图 17 所示, 其中实线对应约束 4 的内容, 虚线对应约束 5 的内容.

在上述约束规则的基础上, 定义逐层反向传播的 $\alpha\beta$ 规则, 如下:

$$R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \left(\alpha \frac{z_{jk}^+}{z_k^+} - \beta \frac{z_{jk}^-}{z_k^-} \right) \tag{18}$$

式中, $z_k^+ = \sum_j z_{jk}^+ + b_k^+ = \sum_j a_j w_{jk}^+ + b_k^+$, $z_k^- = \sum_j z_{jk}^- + b_k^- = \sum_j a_j w_{jk}^- + b_k^-$, 上标 $+$ 和 $-$ 分别表示正和负部分. α 和 β 满足 $\alpha - \beta = 1$. 对上式两边同时求和, 可得:

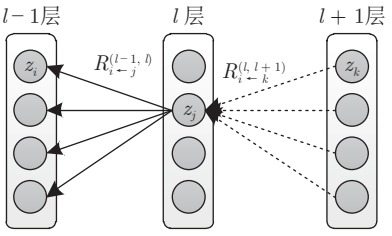


图 17 LRP 反向传播的过程[19]

Fig. 17 The backpropagation process of the LRP[19]

$$\begin{aligned}
\sum_{j=1}^{V(l)} R_{j\leftarrow k}^{(l,l+1)} &= \sum_{j=1}^{V(l)} R_k^{(l+1)} \left(\alpha \frac{z_{jk}^+}{z_k^+} - \beta \frac{z_{jk}^-}{z_k^-} \right) = \\
& R_k^{(l+1)} \left(\alpha \frac{\sum\limits_{j=1}^{V(l)} z_{jk}^+}{z_k^+} - \beta \frac{\sum\limits_{j=1}^{V(l)} z_{jk}^-}{z_k^-} \right) = \\
& R_k^{(l+1)} \left(\alpha \frac{z_k^+ - b_k^+}{z_k^+} - \beta \frac{z_k^- - b_k^-}{z_k^-} \right) = \\
& R_k^{(l+1)} \left(1 - \alpha \frac{b_k^+}{z_k^+} + \beta \frac{b_k^-}{z_k^-} \right)
\end{aligned} \tag{19}$$

以上是化简后的 α和β 规则. 在实际应用中, 取 $\alpha = 1$, $\beta = 0$, 由式 (18) 和式 (19) 分别得到:

$$R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \cdot \frac{z_{jk}^+}{z_k^+} \tag{20}$$

$$\sum_{j=1}^{V(l)} R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \left(1 - \frac{b_k^+}{z_k^+} \right) \tag{21}$$

将 $z_k^+ = \sum_j a_j w_{jk}^+ + b_k^+ = \sum_j z_{jk}^+ + b_k^+$ 代入式 (20) 和式 (21), 并取 $b = 0$, 进一步得到:

$$R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \cdot \frac{a_j w_{jk}^+}{\sum\limits_{j=1}^{V(l)} a_j w_{jk}^+} \tag{22}$$

$$\sum_{j=1}^{V(l)} R_{j\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \tag{23}$$

式 (22) ~ (23) 表示 $\alpha = 1$、$\beta = 0$、$b = 0$ 情形下的相关性传递规则. 相关性值传递与正向过程中的权重和激活值有关, 它们决定了每个神经元的相关性分配比例.

尽管 LRP 的传播规则非常细致, 但初始情形下的相关性值仅考虑了目标类别的神经元, 忽视了其他类别神经元的作用, 即:

$$R_n^{(L)} = \begin{cases} z_t^{(L)}, & n = t \\ 0, & \text{否则} \end{cases} \tag{24}$$

式中, $z_t^{(L)}$ 表示第 L 层目标神经元 t 对应的分值 (Softmax 之前). 该问题导致 LRP 对特定类别的特征不敏感, 即可视化结果不具有类别区分性. 因此, 为了实现类别区分性的可视化, Gu 等[59] 提出了对比式层级相关性反馈 (Contrastive LRP, CLRP). CLRP 区分了初始相关性分数中的目标类和非目标类的比例. 假设 t 表示目标节点, CLRP 将目标节点与其他节点的分数对立, 且其他所有节点的分数固

定, 均为 $-z_t^{(L)}/\ (N-1)$.

$$R_n^{(L)}=\begin{cases}z_t^{(L)}, & n=t\\ -\dfrac{z_t^{(L)}}{N-1}, & \text{否则}\end{cases}\tag{25}$$

式中, N 表示第 L 层神经元的总个数. 进一步地, Iwana 等[60] 提出了 Softmax 梯度层级相关性反馈 (Softmax-gradient LRP, SGLRP), 将 Softmax 梯度信息引入初始相关性分数值, 以区分其中的各个类的比例:

$$R_n^{(L)}=\frac{\partial\hat{y}_t}{\partial z_n}=\begin{cases}\hat{y}_t(1-\hat{y}_t), & n=t\\ -\hat{y}_t y_n, & \text{否则}\end{cases}\tag{26}$$

式中, $\hat{y}_n$ 表示输出层类别 n 的 Softmax 分值, $\hat{y}_t$ 表示输出层目标神经元 t 的分值. 该式将目标节点的相关性分数初始化为偏导数, 其他节点的分数也与其自身的偏导数相关, 但互不相同, 与 CLRP 相比更加合理.

2.1.2.4　深度泰勒分解

深度泰勒分解 (Deep Taylor decomposition, DTD). Montavon 等[61] 认为梯度方法属于 DTD 方法的一种特殊情形, 即用输入空间所有像素的梯度之和表示输入层的相关性分数总和. DTD 的思想基于 CNN 的函数特性, 采用数学中的泰勒分解方法对 CNN 进行分解. 假设将 CNN 视作一个由输入到输出的函数 f, 其中 x 为输入变量, $f(x)$ 为输出结果, 则可以对该函数在某点 a 处进行泰勒分解, 如下:

$$f(x)=f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\cdots+\mathrm{O}((x-a)^n)\tag{27}$$

从数学的角度观察, 右侧第 1 项 $f(a)$ 表示 a 点处的函数值, 右侧第 2 项 $(f'(a)/1!)(x-a)$ 表示该函数的一阶导数 (斜率) 和自变量与点 a 差值的乘积, 右侧第 3 项与函数的二阶导数 (曲率) 相关, 后续依次为更高阶导数项. 而从神经网络的角度观察, $f(x)$ 表示网络输出值, 也即对输入变量 x 的分类结果, 将其分解到输入变量上, 即表示输入变量的贡献值. 假如 a 点为零点, 满足 $f(a)=0$, 则 $f(x)$ 的一阶泰勒项 $f'(a)\cdot x$ 即与 VBP 方法相似, 使用梯度与输入变量的乘积作为相关性值.

从热力图的计算方式看, 梯度方法对应的显著图由 $f'(a)$ 得到, 表示预测结果对输入图像中哪些特征的改变较为灵敏. DTD 对应的显著图则由 $f'(a)\cdot x$ 得到, 显著图中将会包含输入图像中的原始特征, 相当于梯度值与输入图像共同作用的结果. 因此, 与梯度方法相比, DTD 回答了 “是哪些特征让这张图像被分类为猫?”, 而梯度方法则回答了 “是哪些特征让这张图像更像一只猫?”. 这看起来是大致相同的问题, 但实际上前者寻找的是猫的所有特征, 而后者则更偏向于寻找能使猫区分于其他事物的特征, 这也可作为梯度方法为何称作灵敏度分析方法的一种理解[19].

2.1.2.5　小结

本节介绍 4 种基于反向传播的可视化方法, 核心思想是通过设计一定的反向传播规则, 将 CNN 网络输出结果反向传播到输入空间, 为输入空间的各个变量分配相关性值, 以衡量其对预测值的贡献, 这样在输入空间形成了由相关性值构成的图像, 即热力图. 热力图以不同明暗强度的形式表明各像素点与预测值之间的相关性. 最简单的反向传播规则依靠网络自身梯度反向传播中的链式法则, 如 VBP 方法. 更复杂的则采用自定义的反向传播规则, 如 LRP 和 DTD.

2.1.3　类激活映射

1) 类激活映射 (Class activation mapping, CAM) 通过生成类激活图来可视化 CNN 的关注区域, 类激活图使用区域级的特征高亮方式, 以突出与特定类别最相关的区域. Zhou 等[62] 认为, 随着 CNN 层数的加深, 中间层特征图编码中与决策无关的信息越来越少, 因此越往深层目标信息越抽象, 语义信息也越丰富. CNN 最后的卷积层在高层语义信息上达到最佳, 其对应的特征图含有最抽象的目标级语义信息, 且每个通道检测到目标的不同激活部位. 因此, 通过对最后的特征图进行通道级加权调整, 可生成与特定类别最相关的类激活图.

CAM 所依赖的网络结构如图 18 所示, 假设最高层特征图 (第 L 层) 为 $\boldsymbol{A}^L$, 其中, 第 k 个通道为 $\boldsymbol{A}_k^L$, 经过全局平均池化 (Global average pooling, GAP) 层映射到 Softmax 输出层进行分类. 生成类激活图的形式化表示如下:

$$L_{CAM}^c=\mathrm{ReLU}\left(\sum_{k=1}^{n}w_k^c\boldsymbol{A}_k^L\right)\tag{28}$$

式中, w_k^c 表示 Softmax 层第 c 个类别神经元的连接权重. 类激活图的类别信息来源于各通道的权重, 这些权重是与特定类别相关的 Softmax 层权重, 因而合成的是含有类别区分信息的热力图.

2) 梯度加权的类激活映射 (Gradient-weighted CAM, Grad-CAM). 由于 CAM 使用的 GAP 层并

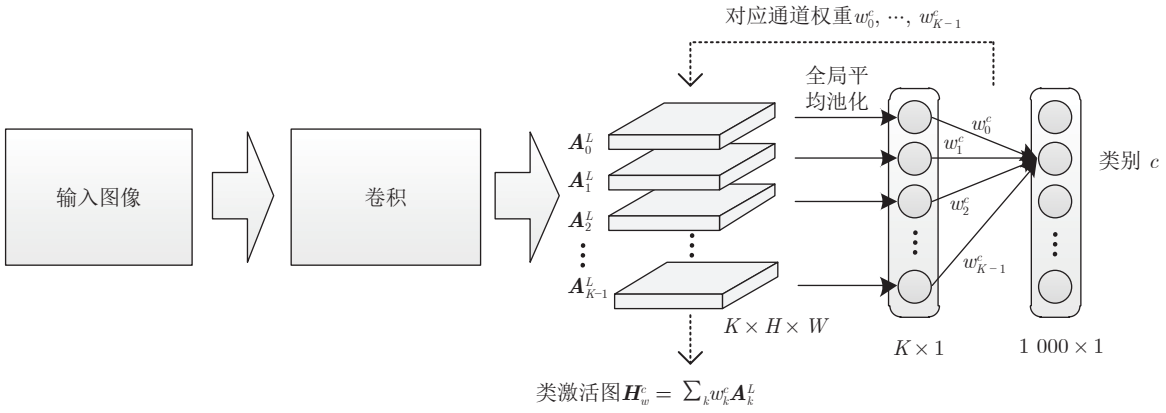


图 18 CAM 的过程

Fig. 18 The process of the CAM

没有出现在 AlexNet[1]、VGGNet[15]和 GooLeNet[14]等常见的网络中, 因此, 若要使用 CAM 可视化 CNN, 需要按照图 18 对网络结构进行改造并重新训练模型, 这极大地增加了工作量. 为了更一般化 CAM, Selvaraju 等[63−64] 提出了基于梯度的 CAM–Grad-CAM. 其基本过程与 CAM 相似, 但为了克服对于 GAP 层的依赖, Grad-CAM 使用反向传播中获取的通道梯度均值作为通道权重, 生成的热力图有类似的效果, 具体过程如图 19 所示.

$$L_{Grad-CAM}^c = \mathrm{ReLU}\left(\sum_{k=1}^{n} \alpha_k^c \boldsymbol{A}_k^L\right) \tag{29}$$

式中, α_k^c 表示各个通道的权重, 计算公式为:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c(\boldsymbol{x})}{\partial \boldsymbol{A}_{k,i,j}^L} \tag{30}$$

式中, 求和元素为通道 k 内部每个神经元激活值的梯度, Z 表示归一化因子. 将得到的类激活图和 GBP 得到的显著图逐点相乘, 能够实现这两种热力图特点的融合, 生成细粒度的、含有类别区分性的热力图.

1) Grad-CAM++. Chattopadhyay 等[65] 进一步提出了 Grad-CAM 的更高阶导数版本 Grad-CAM++, 其基本形式与 Grad-CAM 相同, 仅使用的通道权重不同. Grad-CAM++将更高阶的梯度 (高阶导数) 的组合作为通道权重, 改善了多目标图像的可视化效果, 具体细节这里不再赘述.

2) 分数加权的类激活映射 (Score-weighted CAM, Score-CAM). Wang 等[66] 认为 Grad-CAM 和 Grad-CAM++都是基于梯度的类激活图生成方法, 采用的线性加权权重是由梯度或其变体构成. 而梯度具有不稳定特性[51−52], 将导致生成的类激活图也不稳定. 因此, Score-CAM 试图使用非梯度方法获取各通道的权重, 以消除梯度不稳定性的影响. 如图 20 所示, Score-CAM 将最高层特征图的每个通道作为一个掩码, 将其与输入图像叠加后再送入

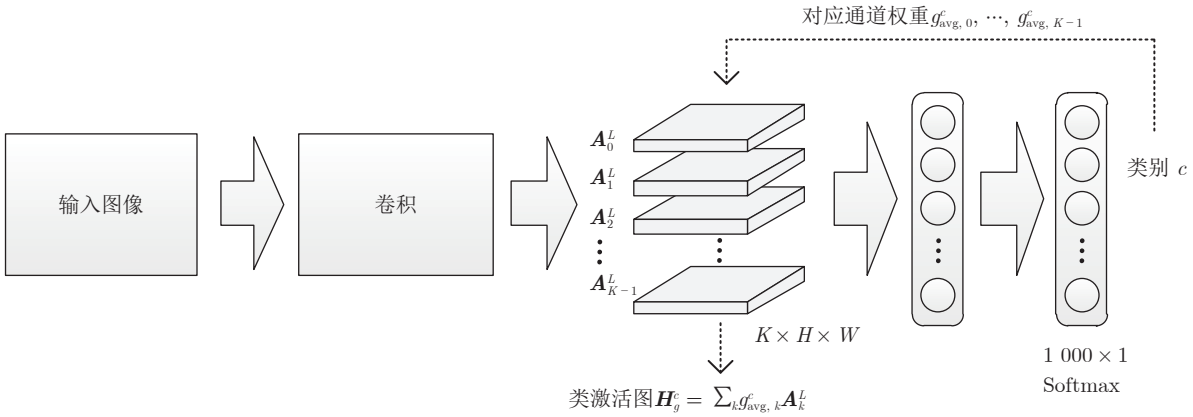


图 19 Grad-CAM 的过程

Fig. 19 The process of the Grad-CAM

CNN 中, 由 CNN 对该通道的各类别的重要性 (即分类概率) 进行预测. 使用特定目标类别的重要性进行通道加权, 可以生成噪声含量更少的类激活图, 并且抵抗一些基于梯度的对抗性攻击, 实现稳定的解释效果.

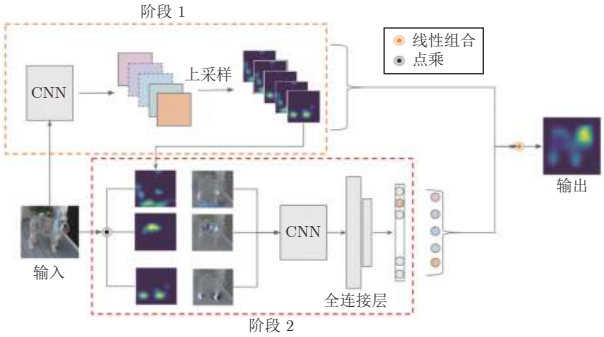


图 20　Score-CAM 的过程[66]

Fig. 20　The process of the Score-CAM[66]

此外, 还有 U-CAM[67] 和 Smooth Grad-CAM++[68] 等以 CAM 作为基础的改进研究. 根据上述几种方法的分析, 可知类激活图的生成过程基本相同, 将最高层特征图作为基本的特征空间, 使用权重对各通道进行加权. 理论上, 权重大小应当表示该通道对特定类别的贡献大小. 经过加权调和后, 将各个通道相加合并, 得到二维的初始类激活图. 此时, 类激活图的尺寸与最高层特征图的尺寸相同, 需采用插值方式 (如双线性插值) 将其扩大到与输入图像相同的尺寸. 在类激活映射基本形式下, 各种方法采用不同的权重对各通道进行加权, 这些权重及其优缺点分析如表 2 所示.

2.1.4　激活最大化

激活最大化 (Activation maximization, AM) 方法用于可视化网络的偏好输入, 找到能够最大限度激活某个特定神经元的输入模式. 与梯度方法的不同之处在于, 梯度方法研究输入数据点通过怎样的变化可使输出分数更高, 这种变化量即作为该点相对于输出的贡献度大小. AM 方法则研究一组怎样的输入数据点集合能够使某个输出类别分数最高, 通过优化方法来搜索这样一组数据点, 即为网络最偏好的某个类别数据点集合. 如图 21 所示, 对于图像分类模型, 使用一张随机初始化的图像, 通过最大化分类该图像的某个激活值, 来反向传播并更新输入图像的像素值, 经过多次迭代, 得到能最大激活该神经元的偏好图像. 该过程中, 分类器的参数保持不变, 仅更新输入图像.

首先, 使用带标注训练集样本 $\{\boldsymbol{x}_t, \boldsymbol{y}_t\}$, 按照设计的目标函数, 对 CNN 进行参数优化并保存, 将该网络称作目标 CNN, 即图中实线部分. 然后, 对随机初始化的输入图像 $\boldsymbol{x}$, 在固定目标 CNN 模型参数的情况下, 通过优化 $\boldsymbol{x}$ 来实现对输出层某个类的概率最大化, 即图中虚线部分. 此时, 以最大化网络输出概率分布中某个类别的概率为目标, 设计目标函数如下[19]:

$$loss = \max_{\boldsymbol{x}}(\ln(p(w_c|\boldsymbol{x}, \theta)) - \lambda||\boldsymbol{x}||^2) \tag{31}$$

该目标函数通过添加对 $\boldsymbol{x}$ 的 L2 正则化约束, 保证其数值变化的稳定性. 使用梯度上升优化算法来更新参数, 此时目标函数的优化对象并不是网络参数, 而是输入矩阵 $\boldsymbol{x}$. 在经过多轮迭代优化后, 最终可得到最大化的类别概率和对应的输入 $\boldsymbol{x}$, 即为该类别神经元所偏好的最佳输入模式. 同时, 目标函数也可以选择最大化中间层某个神经元或一组神经元的激活值, 从而得到中间层神经元偏好的输入模式.

然而, 试图从随机初始化的输入变量中优化出真正与目标类别 c 对应的视觉概念是非常困难的, 实验中经常观察到, 虽然指定类别的概率 $p(w_c|\boldsymbol{x}, \theta)$ 达到迭代停止时的最大值, 但此时对应的输入图像 $\boldsymbol{x}$ 在视觉上却没有任何语义概念. 这表明, 对于恢复出一张有意义的图像来说, 仅依靠类别概率作为先验, 可用的信息量远远不够. 文献 [69] 指出了该问题, 并提出添加一些限制措施 (如正则化约束) 进行改进, 以形成视觉上接近真实图像的生成图像[70−71]. 文献 [19] 从为该优化过程的起点引入先验知识入手, 使用生成网络改进 AM 方法, 提出了基于深度生成模型的 AM 方法 (Deep generator network based AM, DGN-AM), 如图 22 所示.

图 22 所示的过程分为 3 个步骤:

1) 训练目标 CNN. 目标 CNN 表示需要被解释

表 2　类激活映射方法的比较

Table 2　Comparision of the class activation mapping methods

方法	通道权重	优点	缺点
CAM	Softmax 层权重	类别区分性	依赖 GAP 层
Grad-CAM	各通道的梯度平均值	类别区分性, 结构通用	梯度不稳定
Grad-CAM++	各通道的梯度平均值, 高阶梯度	类别区分性, 结构通用	梯度不稳定, 高阶梯度计算复杂
Score-CAM	对各通道的预测值	类别区分性, 结构通用, 权重稳定	权重计算过程复杂, 重复迭代耗时

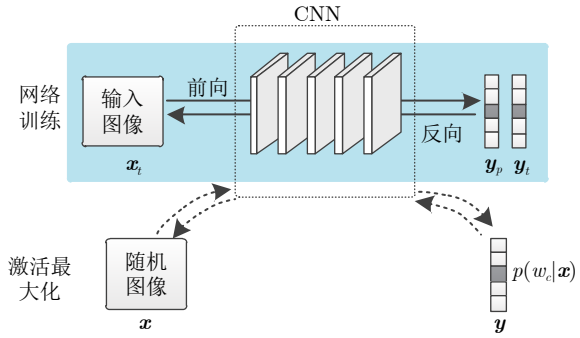


图 21 AM 的过程

Fig. 21 The process of the AM

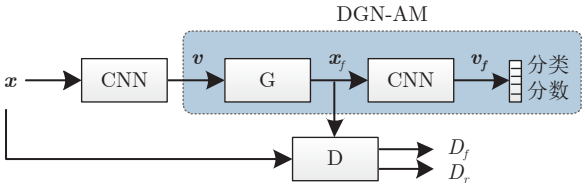


图 22 DGN-AM 的过程

Fig. 22 The process of the DGN-AM

的 CNN, 同时也用来提供先验知识, 作为编码器对输入图像进行编码. 此时, 训练数据来自标准的带标注数据集, 使用交叉熵作为损失函数, 优化参数为目标 CNN 的参数.

2) 训练 GAN. 分别训练生成器 G 和判别器 D, 此时, 固定目标 CNN 的参数, 仅输入无标签的训练数据, 先后对 D 和 G 进行参数优化.

其中, D 的损失函数含有两部分: 最大化 D_r 和最小化 D_f. D_r 表示 D 对真实图像的分类得分, D_f 表示 D 对生成图像的分类得分, 如下:

$$loss_{\rm D} = {\rm E}_x \ln D_f + {\rm E}_x \ln(1 - D_r) \tag{32}$$

G 的损失函数含有最大化 D_f、最小化 $\boldsymbol{x}_f$ 和输入图像 $\boldsymbol{x}$ 之间的欧氏距离, 最小化 $\boldsymbol{v}_f$ (fake) 和 $\boldsymbol{v}$ 之间的距离:

$$loss_{\rm G} = {\rm E}_x \ln(1 - D_f) + \lambda_1 ||\boldsymbol{x} - \boldsymbol{x}_f||^2 + \lambda_2 ||\boldsymbol{v} - \boldsymbol{v}_f||^2 \tag{33}$$

式中, $\boldsymbol{v}$ 表示 $\boldsymbol{x}$ 经过目标 CNN 提取的特征向量, 该向量可作为 $\boldsymbol{x}$ 的代表. $\boldsymbol{x}_f$ 表示 G 作用于向量 $\boldsymbol{v}$ 而生成的图像. $\boldsymbol{v}_f$ 表示使用目标 CNN 从生成图像提取的特征向量. λ_1 和 λ_2 表示损失函数的系数. 该式同时对输入向量及特征向量两部分进行约束, 实现 G 生成图像质量的提升.

由于不需要使用训练数据的标签, 因此上述 GAN 的训练是无监督训练过程. 对 GAN 训练的目的是学习训练样本的先验分布, 从而在后续使用 G 生成图像时, 能够直接使用数据集的先验知识, 而非在完全随机初始化的输入上生成图像.

3) 使用 AM 生成最佳输入模式, 如图 22 虚线框内所示. 针对随机初始化的向量 $\boldsymbol{v}_r$, 经过 G 生成相应的输入图像, 再使用 CNN 分类模型 (即步骤 1 中目标 CNN, 这里用作分类器) 进行分类, 并对输出概率分布中某类的概率进行最大化. 因此, 损失函数与正常的 CNN 分类损失函数相同, 优化对象为输入向量 $\boldsymbol{v}_r$, 其余网络参数均保持不变. 最终得到优化向量, 再将该向量经过 G, 即可生成相应的图像. 由于步骤 2 中训练后的 G 含有丰富的来自于原始训练数据的样本信息, 相当于从目标 CNN 中提取而来, 此时生成图像的视觉效果比单纯使用 AM 时的结果更佳.

如图 23 所示, 表示在 MNIST 数据集上使用 AM 方法对目标 CNN 模型的可视化结果[19]. 其中, 第 1 行表示最简单的情形, 损失函数为 $loss = \max_{\boldsymbol{x}} \ln(p(w_c|\boldsymbol{x}, \theta))$. 第 2 行表示对 $\boldsymbol{x}$ 进行 L2 正则化约束, 损失函数为 $loss = \max_{\boldsymbol{x}}(\ln(p(w_c|\boldsymbol{x}, \theta)) - \lambda||\boldsymbol{x}||^2)$. 第 3 行表示使用数据集的均值 $\boldsymbol{x}_{mean}$ 对 $\boldsymbol{x}$ 进行约束, 损失函数为 $loss = \max_{\boldsymbol{x}}(\ln(p(w_c|\boldsymbol{x}, \theta)) - \lambda||\boldsymbol{x} - \boldsymbol{x}_{mean}||^2)$. 第 4 行表示使用 DGN-AM 方法得到的结果. 可以看出, 随着先验知识的加入, AM 方法针对模型的各个类别提取出的图像越来越逼近真实图像, 逐渐变得清晰可理解. 其中, DGN-AM 方法得到的结果最接近真实图像.

2.1.5 注意力掩码

注意力掩码是由图像识别中的注意力机制产生的掩码矩阵. 由于注意力机制本身能够解释网络对不同变量的依赖, 因此, 对注意力掩码的可视化能够观察到网络对变量的关注度, 从而理解网络的训练效果及决策所依据的输入特征. 图像识别模型中的注意力机制包含通道注意力、空间注意力与通道注意力的混合、类别注意力等.

1) SENet: 2018 年, Hu 等[18] 提出 SENet 模块, 对 CNN 中间层特征图的各通道进行加权调整, 提升特征学习效果. 如图 24 所示, 对于一个尺寸为 $C \times H \times W$ 的多通道特征图, C 表示通道数, H 和 W 分别表示各通道的高和宽, SENet 模块采用压缩和激励 2 个操作, 对特征图进行变换, 得到一个 C 维向量. 使用该向量对各通道进行通道级加权, 实现对通道间依赖关系的显式建模和重标定. 其中, 压缩操作使用全局池化将特征通道池化为单个点, 则整个特征图变为一个特征向量, 该向量称之为全局信息向量, 其每一维均包含对应的整个特征通道信息. 然后, 经过两个全连接层的先降维再还原变

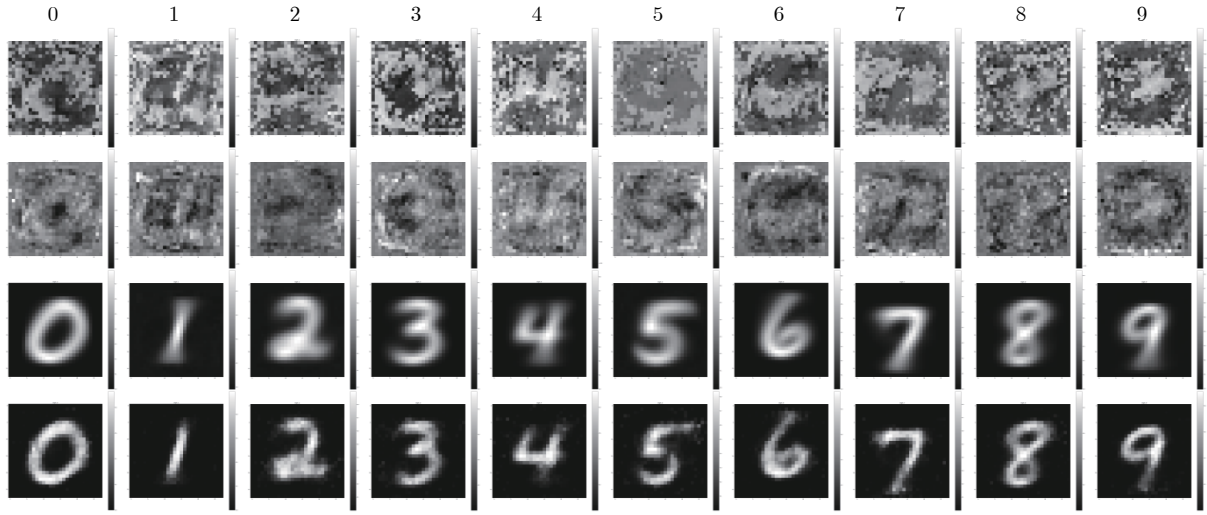


图 23　在 MNIST 数据集上使用 AM 方法对目标 CNN 模型的可视化结果对比[19]

Fig. 23　Comparison of the visualization results of the target CNN model using the AM method on the MNIST dataset[19]

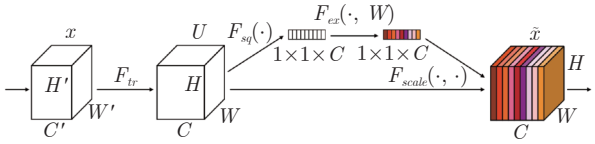


图 24　Squeeze and excitation 模块[18]

Fig. 24　Squeeze and excitation module[18]

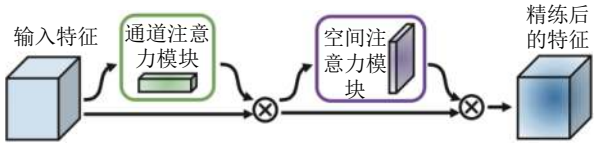


图 25　通道–空间注意力模块[72]

Fig. 25　Channel-spatial attention module[72]

换, 实现自适应校准 (即激励操作), 向量尺寸变化为 $C \to (C/r) \to C$, 其中 r 表示首个全连接层对 C 维向量的降维率. 最后, 将得到的向量作为通道的权重向量, 其每一维表示对应特征通道的权重, 经过通道加权调整后的特征图更具有针对性, 含有更明确的语义信息. 以该方法实现的 SENet 网络, 在 2017 年 ILSVRC 图像分类挑战赛上获得冠军, 实现了 2.25% 的前 5 分类错误率.

2) CBAM (Convolutional block attention module). 2018 年, Woo 等[72] 在通道注意力模块的基础上作了进一步探索, 认为通道级注意力将各通道视为一个整体, 仅能实现各通道间的相互关系建模, 而忽略了通道内不同空间位置的特征分布特性. 为此, 提出了一种通道注意力与空间注意力结合的方式, 对特征图分别进行通道级和空间级的权重调整, 得到通道注意力权重与空间注意力权重, 再与原始特征图融合, 实现对特征图各通道和通道内不同位置的特征重标定. 如图 25 所示. 将通道与空间注意力模块插入到初始网络结构中, 采用先通道后空间注意力模块的顺序, 对特征图进行加权调整, 且不改变特征图尺寸. 与 SENet 相比, CBAM 在各通道之间和通道内各空间位置上实现了更加深入的特征图自适应调整.

通道注意力模块如图 26 所示, 其整体结构与 SENet 大致相似, 但在某些策略上有细微调整, 例如同时使用了使用了最大值池化和平均值池化 2 种池化方式.

通道注意力权重由各通道全局池化得到, 因此通道注意力权重向量的维度与通道数相同. 与之不同, 空间注意力权重按照各空间位置进行逐通道池化, 因此空间注意力权重为矩阵形式, 其尺寸与各通道尺寸大小相同, 空间注意力模块如图 27 所示.

3) 类别注意力. 对于一些具有类别区分信息的可视化方法, 如 CAM 和 Grad-CAM 等, 常被集成到网络结构中, 通过生成针对特定目标的注意力掩码, 引导网络将更多注意力应用到这些区域的特征提取和语义表示上, 以针对性地提升模型的学习性能. 文献 [73] 提出使用一个额外的预训练模型辅助生成 CAM 类激活图, 将该图二值化后作为掩码添加在原图上, 从而集中对原图中目标主体区域的关注. 文献 [74] 同样将该思路应用到图像分类任务中. 文献 [75] 使用自顶向下的注意力将类别信息传递到特征空间, 实现针对特定类的注意力. 文献 [76] 利用指定类的注意力实现物体级别的区域关注, 从而对多标签图像进行分类等. 文献 [77] 使用 Grad-CAM 为遥感图像的多标签分类任务提供类别注意

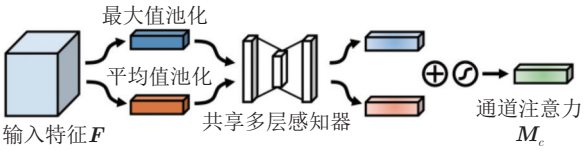


图 26 通道注意力模块[72]

Fig. 26 Channel attention module[72]

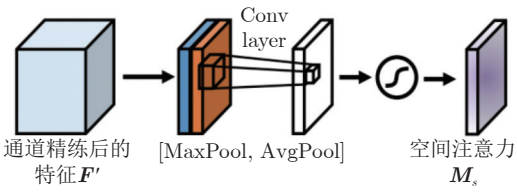


图 27 空间注意力模块[72]

Fig. 27 Spatial attention module[72]

力等. 这些研究在不同任务中应用类别注意力, 而非聚焦于可视化方法本身, 可作为表征可视化方法的拓展研究.

基于注意力掩码的可视化呈现出以下特点: 1) 注意力可视化多用于诊断网络的训练效果, 通过观察内部表征来推测注意力效果的好坏, 但无法对 CNN 的决策依据进行解释. 2) 网络分类/检测准确率越高, 中间层注意力掩码对目标的信息覆盖越全面, 则特征图对目标的定位越准确. 图 28 展示了在 ResNet50 中使用 SENet 和 CBAM 两种模块的注意力效果对比.

2.1.6 其他方法

除了上述几类方法外, 还有一些方法在图像和文本任务的解释上均适用, 常被用于特征归因及特征选择, CNN 可视化仅是其中一项应用. 本节介绍 2 种常见的方法: 局部可理解的模型无关解释 (Local interpretable model-agnostic explanations, LIME) 和沙普利加和解释 (Shapley additive explanations, SHAP).

1) Ribeiro 等[78] 于 2016 年提出 LIME 方法, 用于解释任意黑盒分类器的预测. LIME 方法的主要思想为, 在输入样本附近多次采样, 获取一组近邻样本, 使用这些近邻样本训练可解释的线性模型, 利用线性模型在局部范围内逼近深度模型的预测, 实现模型代理解释. LIME 可应用于图像和文本等分类器的解释.

具体地, 给定输入样本 x, 目标分类器 f, 定义解释模型 g, LIME 通过不断在输入样本的邻域内采样并作为输入, 来优化下列目标函数:

$$\xi(x) = \arg\min_{g \in G} loss(f, g, \pi_x) + \Omega(g) \tag{34}$$

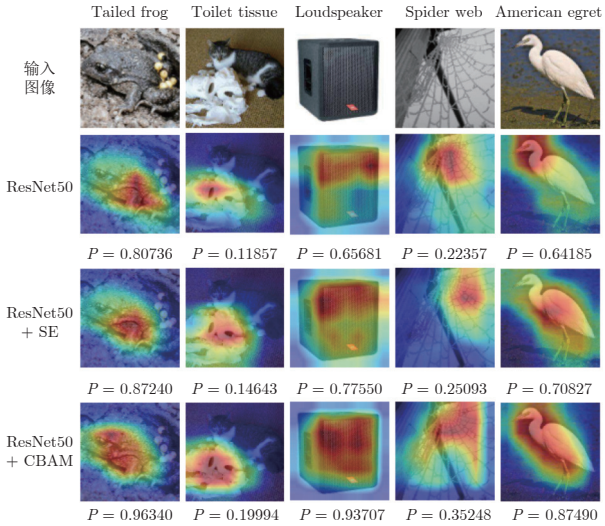


图 28 ResNet50、集成 SENet 的 ResNet50 (ResNet50 + SE) 和集成 CBAM 的 ResNet50 (ResNet50 + CBAM) 的最高层特征图的可视化[72]

Fig. 28 Visualization of the highest-level feature maps of ResNet50, ResNet50 integrated with SEnet (ResNet50 + SE), and ResNet50 integrated with CBAM (ResNet50 + CBAM)[72]

式中, π_x 表示采样样本与输入样本之间的相似性度量, $\Omega(g)$ 约束解释模型的复杂度, G 表示一组可解释的简单模型. 该式的前一项用于保证解释模型的保真度, 使其在局部的预测结果与目标分类器尽量一致. 第 2 项用于约束解释模型的复杂度. 相似性度量函数用于衡量扰动前后样本的相似度, 这个相似度即可用作采样样本的权重:

$$\pi_x(z) = \mathrm{e}^{-\frac{D^2(x,\,z)}{\sigma^2}} \tag{35}$$

式中, $D(\cdot)$ 表示距离函数, 如 L2 范数距离. 将 LIME 实例化为对某个图像分类器的解释, 其中的训练样本构建流程见图 29.

图 29 中, x' (d' 维, 可理解的特征) 表示输入样本 x (d 维, 原始特征) 的可理解表示形式, 例如二值向量, 每一维表示 x 的某个特征出现与否. 将这种映射记作 $x = h_x(x')$. 在 x' 附近随机采样 N 次, 得到 N 个样本 z_i'. 将 z_i' 恢复到原始输入空间, 得到 z_i, 计算相应标签 $f(z_i)$ 及与 x 的相似度 π_x. 其中, 采样的过程就是在可理解的表示域内扰动, 扰动的对象为一组相邻的超像素 (即一系列相邻且相似的像素形成的像素块), 通过改变 $x' \in \{0, 1\}$ 的值, 来表示对应超像素的出现与否, 从而获取采样样本.

按照上述流程, 获得采样样本及其标签, 形成输入样本 x 的近邻数据集 $Z = \{z_i', f(z_i)\}_N$, 可在 Z 上优化如下目标函数, 得到解释模型 g:

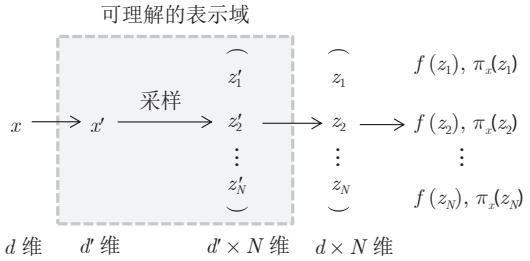


图 29　LIME 的样本处理流程

Fig. 29　The sample processing flow of the LIME

$$loss(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z)(f(z) - g(z'))^2 \tag{36}$$

直观来看, LIME 希望在局部范围内, 使用采样得到的简单样本 z' 训练出解释模型 $g(z')$, 使得在 $z' \approx x'$ 的局部范围内, 总有 $g(z') \approx f(h_x(z'))$. 这样训练得到的解释模型 g 便可以在局部代替待解释的目标模型 f.

如图 30 所示, 表示 LIME 在 3 种不同的 CNN 模型上针对输入图像的可视化结果, 此时设置随机采样样本数量为 10 000, 使用余弦函数作为距离度量. 该方法能够对特定预测结果分别找到正向和负向贡献的特征, 且不同模型上的可视化效果也不同. LIME 方法的优点在于简单且易于理解, 是一种模

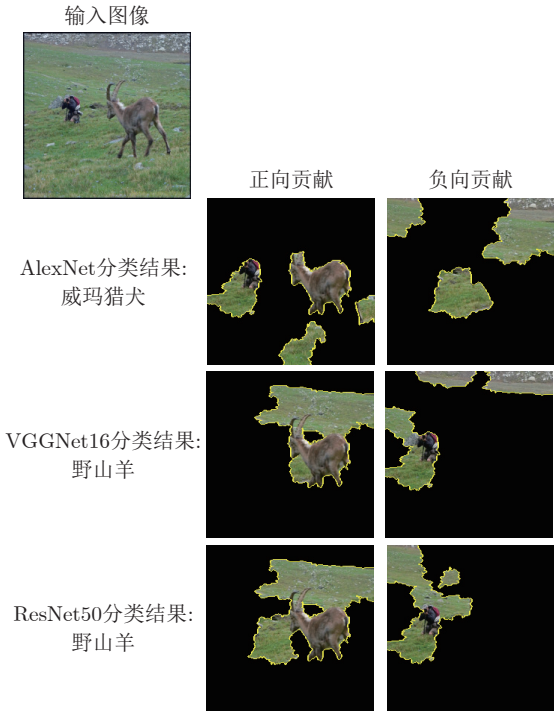


图 30　LIME 在 AlexNet、VGGNet16 及 ResNet50 模型上可视化结果示例

Fig. 30　Example of LIME visualization results on AlexNet, VGGNet16 and ResNet50 models

型无关的方法, 通用性较好, 对于图像和文本分类任务均适用. 但其缺点在于其采用的是局部近似, 无法对模型整体进行解释, 并且需要重新训练一个新的解释模型. 同时, 对于同一个输入样本, 多次运行 LIME 方法时, 由于每次随机采样的样本不同, 使得每次解释结果都不相同, 且所选择的采样样本数量和特征数目等参数都会影响解释效果, 使其无法为用户提供一个稳定的解释.

2) Lundberg 等[79] 利用博弈理论的 Shapley 值来解释模型预测, 提出了 SHAP 方法. Shapley 值由加州大学洛杉矶分校的 Shapley 等[80] 提出, 用于解决合作博弈中的分配均衡问题, 在许多领域均有应用. SHAP 方法的基本思想为, 将输入数据中的特征视为合作博弈模型中的玩家, 通过计算每个玩家的 Shapley 值来量化该玩家在完成任务中的贡献, 进而为其分配相应的贡献值, 即该特征对预测结果的贡献.

假设有目标模型 f 和解释模型 g, 且 f 和 g 满足:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i = f(z) \tag{37}$$

式中, $z'_i \in \{0, 1\}^M$ 表示对应特征出现与否, M 表示输入特征的数量, $\phi_i \in \mathbf{R}$ 表示第 i 个特征的 Shapley 值, ϕ_0 是解释模型的常数. 该解释模型相当于带常数项的二值变量的线性函数, 满足以下 3 点性质:

1) 局部保真性. 即对于同一个输入样本 z, 解释模型 g 的输出 $g(z')$ 应当与目标模型 f 的输出 $f(z)$ 保持一致.

2) 缺失性. 若输入中某个特征没有贡献, 则其 Shapley 值为 0, 即 $z'_i = 0 \rightarrow \phi_i = 0$.

3) 一致性. 如果目标模型发生变化, 使得某个特征的贡献增加, 则其对应的 Shapley 值也应增加.

Shapley 值是已被证明满足有效性、对称性、可加性和单调性的指标. 按照上述性质, 每个特征都具有唯一的 Shapley 值. Lundberg 等[79] 给出了具体的 Shapley 值计算公式. 文献 [79] 还介绍了几种基本的 SHAP 方法变体, 如 Kernel SHAP、Linear SHAP、Low-Order SHAP 和 Deep SHAP. 其中, Kernel SHAP 就是线性模型下的 LIME 方法与 Shapley 值的结合. SHAP 方法的优势在于其广泛适用性, CNN 决策结果的可视化仅是其应用的一个方面, 对于其他机器学习模型的解释也都适用.

2.2　分析与比较

2.2.1　特点分析

第 2.1 节介绍了多种典型的可视化方法, 本节

对它们的特点进行了归纳, 如表 3 所示, 分为以下 6 个方面:

1) 细粒度与区域级. 细粒度的可视化实现像素级的相关性反向传播, 输入空间中每个像素都会被分配一个对应的贡献值, 在数值上表示其对 CNN 某个特定输出值的贡献大小. 基于反向传播的方法多属于细粒度的可视化方法. 区域级的可视化一般不关注单个像素与输出之间的相关性, 而是将某些区域所涵盖的具有语义概念的像素集合作为整体, 其中单个像素的数值大小与其对网络输出值的影响并不一定呈现比例关系, 但区域整体上对 CNN 的输出类别贡献成比例. 基于扰动的方法、类激活映射及注意力掩码等多属于区域级的可视化方法.

2) 类别相关性. 通用的可视化方法生成的热力图是与具体输出类别无关的, 比如梯度方法、反卷积、LRP 和 DTD 等. 输入图像中含有明显语义信息的目标都会被可视化方法定位, 但无法区分哪些目标和当前输出类别最相关. 类别区分性可视化方法生成的热力图是类别相关的, 可在输入空间中找到与指定输出类别相关的区域和特征. CLRP、SCLRP、AM 和 CAM 这类方法均为类别区分性可视化方法.

3) 在线与离线可视化. 离线可视化表示对已经完成训练的模型的可视化, 在可视化过程中仅通过测试样例输入, 而无需对模型本身的结构或参数进行任何修改. 在线可视化则需要修改模型结构并重新训练模型, 典型的在线可视化方法如 CAM, 需要为网络末端添加 GAP 层, 然后重新训练模型.

4) 模型明晰的和模型不可知的. 模型明晰的方法将模型视作白盒, 事先知道模型结构并能获取模型的参数和激活值等内部信息. 模型不可知的方法将模型视作黑盒, 仅能获取模型的输入和输出, 对模型的其他信息 (如网络结构和参数等) 并不了解. 除简单扰动、LIME 和 SHAP 方法外, 其他均为模型明晰的方法.

5) 可视化视角. 包括从解释神经元的激活、层的表征、输出类别等方面来理解 CNN 的表征. 以反向传播可视化为例, 通常以输出类别的分数开始,

表 3 可视化方法的特点比较

Table 3 Comparison of characteristics of visualization methods

方法分类	方法名称		发表年份	细粒度/区域级	类别相关	在线/离线	模型明晰/模型不可知	可视化视角	局部解释/全局解释
扰动	简单扰动[13, 42–43]		2014、2018	区域级	否	离线	模型不可知	输出类	局部
	有意义的扰动[44]		2017	区域级	否	离线	模型明晰的	输出类	局部
	生成式扰动[45–46]		2019	区域级	是	离线	模型明晰	输出类	局部
反向传播	梯度类反向传播	VBP[22–23]	2010、2013	细粒度	否	离线	模型明晰	输出类	局部
		GBP[50]	2014	细粒度	否	离线	模型明晰	输出类	局部
		Smooth gradient[52]	2017	细粒度	否	离线	模型明晰	输出类	局部
		Integrated gradient[53]	2017	细粒度	否	离线	模型明晰	输出类	局部
		Rectified gradient[54]	2019	细粒度	否	离线	模型明晰	输出类	局部
	规则类反向传播	Deconvolution[13]	2013	细粒度	否	离线	模型明晰的	神经元/层	局部
		LRP[58]	2015	细粒度	否	离线	模型明晰	输出类	局部
		DTD[61]	2017	细粒度	否	离线	模型明晰	输出类	局部
		CLRP[59]、SGLRP[60]	2018、2019	细粒度	是	离线	模型明晰	输出类	局部
类激活映射	CAM[62]		2015	区域级	是	在线	模型明晰	输出类	局部
	Grad-CAM[63–64]		2016、2017	区域级	是	离线	模型明晰	输出类	局部
	Grad-CAM++[65]		2018	区域级	是	离线	模型明晰	输出类	局部
	Score-CAM[66]		2019	区域级	是	离线	模型明晰	输出类	局部
激活最大化	AM[81]		2009	细粒度	是	离线	模型明晰	神经元/输出类	全局
	DGN-AM[82]		2016	细粒度	是	离线	模型明晰的	神经元/输出类	全局
注意为掩码	通道注意力[18]		2017	区域级	否	在线	模型明晰的	层	局部
	空间–通道注意力[72]		2018	区域级	否	在线	模型明晰	层	局部
	类别注意力		—	区域级	是	在线	模型明晰	层	—
其他方法	LIME[78]		2016	区域级	是	离线	模型不可知	输出类	局部
	SHAP[79]		2017	细粒度	是	离线	模型不可知	输出类	局部

反馈到输入空间得到热力图, 即表示从输出类的角度进行可视化. 注意力掩码通常以在线的方式, 直接观察中间层的特征图, 从而理解 CNN 模型训练效果的好坏.

6) 局部解释和全局解释. 局部解释一般仅以单个输入样本为指导, 根据对该样本的可视化结果来理解 CNN 在该输入下的表征与决策. 全局解释侧重从整个模型的角度, 理解模型所学习到的知识和决策规则等. 显然, 解释 CNN 对单个样本的决策结果比理解 CNN 模型整体更加简单, 因此, 局部解释比全局解释在方法实现上更加容易, 表 3 所列方法多数属于局部解释方法.

2.2.2　结果比较

由于不同方法生成的热力图效果并不相同, 后处理对于最终结果的比较非常重要, 图 31 总结了一些基本的热力图处理技巧.

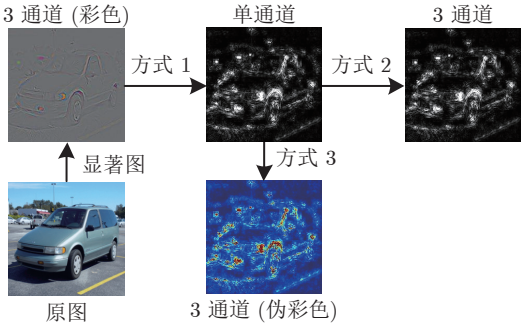


图 31　热力图的后处理与效果对比

Fig. 31　Post-processing and effect comparison of heatmap

方式 1 表示 3 通道热力图 (如 VBP、GBP、反卷积等基于反向传播的方法) 在各通道像素值的相加求和后, 可转换为单通道热力图, 视觉效果上等同于将彩色图转换为灰度图.

方式 2 表示单通道热力图 (如 CAM) 在通道复制后, 可转换为 3 通道热力图. 由于各通道数值相同, 因此, 其视觉效果与单通道热力图相同.

方式 3 表示单通道热力图 (如 CAM) 转换为 3 通道热力图. 与方式 2 不同的是, 该方式使用 OpenCV 等工具包中的伪彩色函数生成伪彩色图. 这种后处理在 CAM 类方法中比较常见, 用于区分图像中的物体类别.

利用上述热力图后处理方法, 选择几种典型可视化方法进行效果对比, 图 32(a) 为 VBP[23], 图 32(b) 为 GBP[50], 图 32(c) 为 Smooth gradient[52], 图 32(d) 为 Integrated gradient[53], 图 32(e) 为 Guided Grad-CAM[63], 图 32(f) 为 LRP[58], 图 32(g) 为 Grad-CAM[63], 图 32(h) 为 Score-CAM[66], 图 32(i) 为简单扰动[13], 图 32(j) 为有意义的扰动[44]. 图 32 对比展示了这些方法的测试效果[83].

从可视化结果来看, 基于梯度的方法生成散点形式的热力图, 这与梯度方法的逐元素反向传播有关. 上述几种梯度方法生成的散点图中的噪声越来越少, 效果上呈现出逐渐改善的趋势. 对于不同的输入图像, 由于图像中前景物体及背景的复杂程度不同, 使用某一种方法得到的可视化效果并不总是特别理想. 但总体上看, GBP 和 Smooth gradient 方法更能将关注度集中在前景物体上. LRP 方法则倾向于可视化目标的轮廓特征, 可作为一种物体边缘检测器使用.

CAM 类方法生成的类激活图能够实现区域级可视化效果, 这一点与梯度类方法区别较为明显. CAM 类方法的优势在于类别区分性, 这种特性使其能够应用于含有多个不同类目标的场景, 实现目标定位任务. 设定适当的阈值二值化处理类激活图, 然后生成目标的边框, 即可实现图像级标签下的弱监督目标定位.

基于扰动的方法对应的热力图较为平滑, 这与所使用的扰动策略有关. 扰动方法需要多次前向传播, 相比于其他方法更耗时. 由于扰动方法直接在输入空间进行修改, 有时可能无意中扰动出对抗性输入, 使相应的可视化结果出现误差[44]. 另外, 若所能获取的有关目标模型的知识有限, 可优先选择扰动方法进行可视化, 因为扰动方法多数是模型不可知的方法, 这是其重要优势.

3　可视化效果的评估

可视化效果评估用于度量不同方法的解释效果, 指导用户针对特定任务选择合适的可视化方法. 具体来讲, 可从以下两个方面对可视化效果进行评估: 有效性和鲁棒性.

3.1　有效性

3.1.1　定性评估

定性评估方法在表征可视化研究的早期被经常使用, 依靠人的视觉感观来评价解释结果是否符合人的认知. 由于定性评估具有简单直观、便于理解等优点, 至今仍广泛使用. 常用的定性度量标准有以下 3 个:

1) 视觉连贯性. 热力图需要关注感兴趣的目标区域, 忽略其他不相关区域. 在视觉连贯性标准下, 热力图中突出的区域对感兴趣目标的覆盖越全面、冗余部分越少, 表明可视化效果越好.

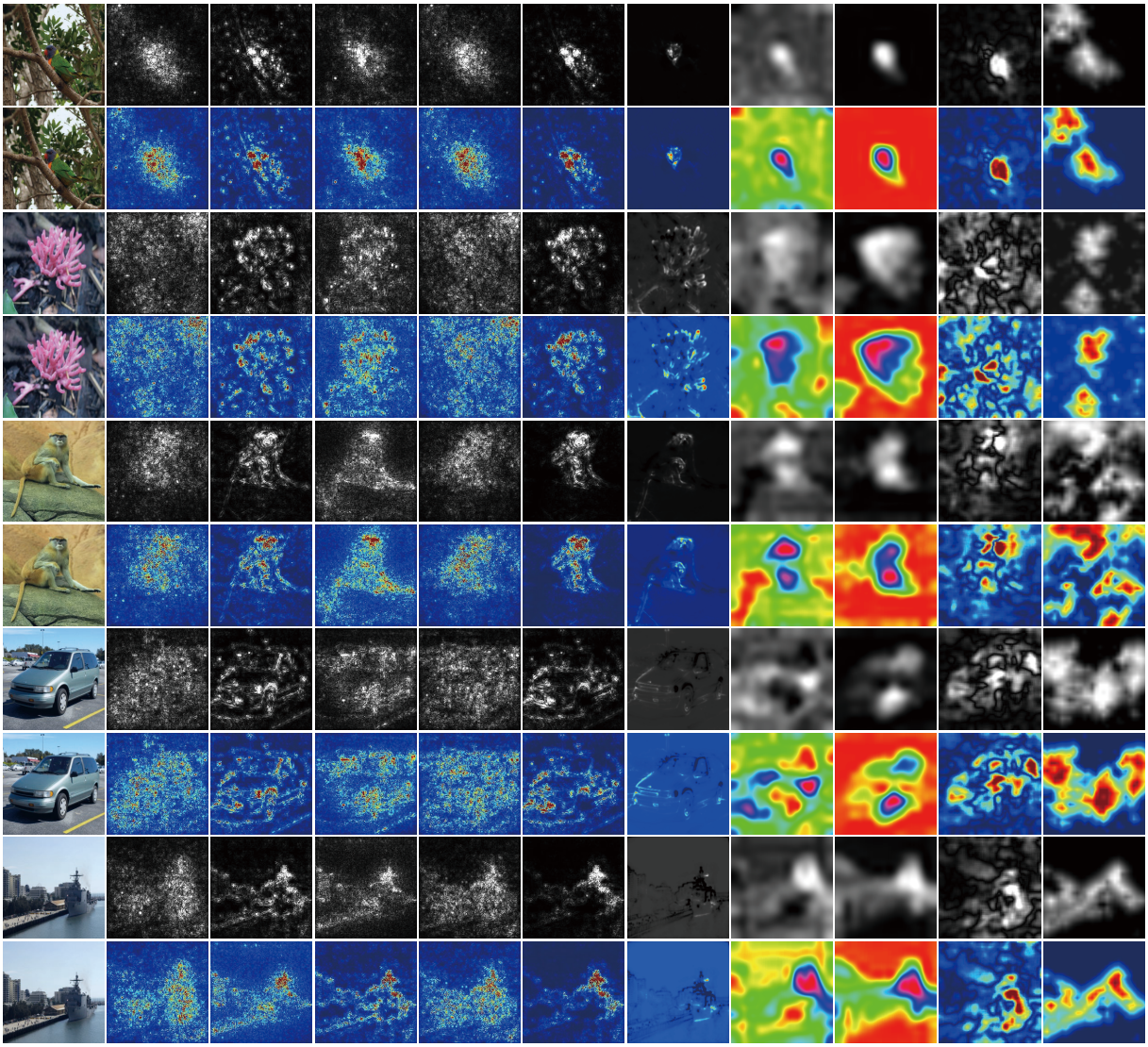


图 32 可视化方法的效果比较. 每张输入图像分别展示了灰度和彩色两种可视化结果

Fig. 32 Comparison of the effects of visualization methods. Each input image shows two visualization results of grayscale and color image

2) 视觉可分辨性. 热力图需要与预测结果中的不同类别相对应, 这对于含有多个不同类别目标的图像来说至关重要. 例如, 在同时含有 “Dog”和 “Cat”的图像中, “Dog”的热力图应该聚焦与之对应的图像区域, 而尽量去除与 “Cat”相关的区域. 视觉可分辨性对应于可视化方法的类别区分性特点, 用于评估热力图能否针对特定类别进行可视化, 以定位仅与该类别其相关的区域和特征.

3) 多目标可视化. 当多个同一类别的目标同时出现在图像中时, 可视化方法需要同时定位多个目标, 而没有遗漏其中的某个目标.

此外, 解释结果的客观性也应作为一种评价可视化方法有效性的标准, 即解释结果是否真实的反映了模型的决策依据, 而非按照 “预期”寻找到了人类所希望的决策依据. 例如, 文献 [78] 的实验表明, 分类器做出的决策可能依据目标周围的环境因素, 而目标自身却不是主导因素. 该情形下, 解释方法只能对分类器的分类依据如实解释, 而非按照人的期望去定位目标主体的某些特征. 文献 [84] 的研究同样验证了这一点, 若分类器从数据集中学习到 “女性面部涂口红, 男性面部不涂口红” 这种带有一定 “偏见”的事实, 导致分类器面对 “涂有口红”的男性图像时, 仍将其判定为 “未涂口红”. 此时, 尽管分类器结果错误, 但解释方法应当遵循分类器的决策依据, 定位于男性面部的其他特征, 将其作为判定为 “男性”, 进而 “未涂口红”的依据. 而非像人所认为

的那样, 直接对该图像的嘴唇位置进行定位, 这样的解释结果与分类结果将出现明显不一致的现象, 无法客观地解释分类器的决策依据.

尽管解释的客观性问题在相关文献中较少被提及, 但也应引起注意. 只有让解释方法客观、真实地反映模型的决策依据, 才能使人真正理解并诊断其存在的问题, 进而改进与优化.

3.1.2 定量评估

定量评估方法按照某种得分规则, 计算数据集上所有图像的可视化结果的平均得分, 从而定量比较各方法的优劣. 这里介绍 3 种典型的定量评估方法.

1) 弱监督目标定位[63−64]. 使用目标定位任务的指标来评价可视化方法的目标定位效果. 具体方法为: 按照设定的阈值处理热力图以生成边框, 然后和真实边框进行比较, 计算交并比 (Intersection over union, IoU). 对于某个定位结果, IoU >0.5 表示成功定位该目标, 以此在整个数据集上计算定位准确率. 该方法多用于评价 CAM 这类目标区分性较好、具有区域级可视化效果的方法.

由于某些细粒度的可视化方法更易定位与预测最相关的像素, 而非寻求覆盖目标整体, 因此, 热力图对应的边框将会定位在目标的局部区域, 导致 IoU 值总体偏小. 此时, IoU 值无法反映解释结果的优劣, 表明这种评价方法具有一定的局限性[66].

2) 指向游戏[75]. 对于特定类别目标的热力图, 计算其最大激活值是否落入该类别的一个实例的边框中, 若落入则计入指向成功 1 次 ($\#\,Hit$), 否则不计入 ($\#\,Miss$), 以此计算每个目标类别的定位准确率 $Acc = \#Hits/(\#Hits + \#Misses)$. 最终使用不同类别的平均准确度作为度量标准.

指向游戏只考虑热力图的最大值点, 无需突出特定目标的全部区域, 仅需对热力图最少量的后处理, 这样对不同特点的热力图更公平. 其可能的缺点在于热力图自身的噪声问题, 最大值点可能来自极值噪声点, 导致评价结果产生误差.

3) 随机性检验. 文献 [85]提出随机性检验方法, 用于评估可视化方法的适用范围和解释质量. 分为两种随机化检验: 一种是模型参数随机化, 使用随机化模型参数和预训练模型参数加载模型, 对比这两种情形下可视化方法的输出变化, 以检验该方法是否对模型参数敏感; 另一种是数据随机化, 对训练数据标签进行随机化打乱并重新训练模型, 与未打乱标签的可视化结果进行对比, 检验该方法是否对训练数据标签敏感.

随机性检验已成为广泛认可的基准测试方法, 用于检验可视化方法是否能有效实现解释, 从而区分出对模型参数和训练数据标签并不敏感的可视化方法. 这种不敏感的可视化方法的真实作用相当于一个独立于模型的边缘检测器, 而非一个有效的解释器. 文献 [85] 通过该实验验证了 VBP 和 Grad-CAM 的有效性, 而 GBP 和 Guided Grad-CAM 等未通过检验.

3.2 鲁棒性

可视化方法的鲁棒性与 CNN 模型的鲁棒性不同. CNN 模型的鲁棒性是指模型的预测结果不会因为对抗攻击而发生明显变化. 可视化方法的鲁棒性是指在面临对抗攻击时, 可视化方法仍能够提供准确有效的解释. 为此, 本文将对抗攻击分为以下 2 种情形: 1) 攻击模型预测结果, 测试解释结果是否随之改变; 2) 攻击解释结果, 测试其是否会被误导.

3.2.1 稳定性

可视化方法的稳定性是指在模型预测受到对抗攻击时, 可视化方法的解释结果仍能保持稳定而不发生显著变化. 其中, 用于攻击模型预测结果的对抗样本 $\boldsymbol{x}_{adv}$ 具有以下 3 个特点:

1) 对原图 $\boldsymbol{x}$ 施加扰动 δ 后得到对抗图像 $\boldsymbol{x}_{adv}$, $\boldsymbol{x}_{adv}$ 相对于 $\boldsymbol{x}$ 的变化在视觉上难以感知, 满足 $||\delta|| = ||\boldsymbol{x}_{adv} - \boldsymbol{x}|| \ll \varepsilon$ (ε 表示较小常数), 保证扰动后图像的视觉不变性;

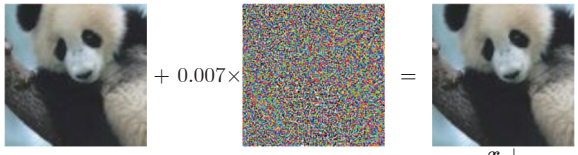
2) 图像分类模型 f 对 $\boldsymbol{x}_{adv}$ 的分类结果将会极大的改变, 即 $f(\boldsymbol{x}_{adv}) \neq f(\boldsymbol{x})$;

3) 解释方法 g 产生的解释结果不会因为扰动而发生显著变化, 满足 $g(\boldsymbol{x}_{adv}) \approx g(\boldsymbol{x})$.

这里简单介绍一种经典的基于梯度的对抗攻击方法 (Fast gradient sign method, FGSM)[86], 可用于攻击模型的预测结果, 检验可视化方法的解释结果是否仍保持稳定. FGSM 利用梯度上升方法, 通过优化输入图像来最大化损失函数, 使模型产生误分类的结果, 此时对应的输入图像即为对抗图像. FGSM 方法形式化如下:

$$\boldsymbol{x}_{adv} = \boldsymbol{x} + \epsilon \cdot \text{sign}\left(\nabla_{\boldsymbol{x}} J\left(\theta, \boldsymbol{x}, f(\boldsymbol{x})\right)\right) \tag{38}$$

式中, $\nabla_{\boldsymbol{x}} J(\theta, \boldsymbol{x}, f(\boldsymbol{x}))$ 表示原图对应初始类别的梯度. sign(.) 表示符号函数, 根据梯度正负取+1 或 −1. ϵ 表示扰动系数. FGSM 的具体过程如图 33 所示. 其中, $\boldsymbol{x}$ 表示输入图像, $f(\boldsymbol{x})$ 的结果为 “Panda”, 置信度为 57.7%. 扰动量大小 $\epsilon = 0.07$. 经过 “Nematode”的扰动后, 扰动后的图像虽然在视觉上仍为 “Panda”, 但却被分类为 “Gibbon”, 且置信度高达 99.3%.



$\boldsymbol{x}$
"Panda"
57.7% 置信度

sign ($\nabla_x J(\theta, \boldsymbol{x}, y)$)
"Nematode"
8.2% 置信度

$\boldsymbol{x}+$
ϵ·sign ($\nabla_x J(\theta, \boldsymbol{x}, y)$)
"Gibbon"
99.3% 置信度

图 33 FGSM 生成对抗样本的过程[87]

Fig. 33 The process of generating adversarial example by FGSM[87]

文献 [63] 和文献 [66] 使用 FGSM 对抗样本测试 Grad-CAM 生成的类激活图的稳定性, 如图 34 所示. 图 34(a) 和 (b) 分别表示原图和对抗图像, 原图分类结果为 Boxer: 0.40, Tiger Cat: 0.18. 对抗图像的分类结果为 Airliner: 0.9999. 在图 34(c) 和图 34(d) 中, 针对对抗图像, 使用 Grad-CAM 分别对 Boxer (Dog) 和 Tiger Cat (Cat) 进行定位时, 仍可以稳定地找出相关目标区域, 尽管此时这两种目标的分类置信度非常低. 这表明 Grad-CAM 产生的解释具有一定的稳定性, 可以抵抗针对模型预测结果的对抗攻击.

Boxer: 0.40
Tiger Cat: 0.18
(a)

Airliner: 0.9999
(b)

Boxer: 1.1×10^{-20}
(c)

Tiger Cat:
6.5×10^{-17}
(d)

图 34 使用 FGSM 对抗样本测试 Grad-CAM 的稳定性[63]
((a) 原图; (b) 对抗图像; (c) Grad-CAM "Dog";
(d) Grad-CAM "Cat")

Fig. 34 Using FGSM adversarial example to test the stability of Grad-CAM[63] ((a) Original image; (b) Adversarial image; (c) Grad-CAM "Dog"; (d) Grad-CAM "Cat")

尽管这是一种测试可视化方法稳定性的方法, 但文献 [88] 认为, 当模型分类结果受到攻击时, 解释结果应当随着分类结果的改变而改变, 即解释方法应该尝试对新的分类结果进行解释, 而不应保持原来的解释不变, 这样才是一种忠实的解释方法. 由此可见, 这种稳定性测试方法的合理性仍存在一定疑问. 根据这种思路, 即可视化结果应当与新的误分类结果相对应, 文献 [88] 使用可视化方法来检测对抗样本, 从而找出其中误导分类结果的特征.

3.2.2 抗欺骗性

可视化方法的抗欺骗性是指可视化方法自身受到对抗攻击时, 解释结果能够抵抗这种欺骗性的攻击, 仍能实现有效的解释.

文献 [89] 指出, 可视化方法生成的显著图可以被人为设计的对抗样本操纵. 通过对输入施加视觉上难以察觉的扰动, 使网络的输出近似保持不变, 显著图却可以被任意改变. 也就是说, 这种对抗样本的攻击对象不是模型预测结果, 而是对预测结果的解释. 用于攻击可视化方法的解释结果的对抗样本 $\boldsymbol{x}_{adv}$ 具有以下 3 个特点[90]:

1) 对原图 $\boldsymbol{x}$ 施加扰动 δ 后得到对抗图像 $\boldsymbol{x}_{adv}$. $\boldsymbol{x}_{adv}$ 相对于 $\boldsymbol{x}$ 的变化在视觉上难以感知, 满足 $||\delta|| = ||\boldsymbol{x}_{adv} - \boldsymbol{x}|| \ll \varepsilon$ (ε 表示较小常数), 保证扰动后图像的视觉不变性;

2) 图像分类模型 f 对 $\boldsymbol{x}_{adv}$ 的分类结果基本不变, 即 $f(\boldsymbol{x}_{adv}) = f(\boldsymbol{x})$;

3) 解释方法 g 产生的解释结果 $g(\boldsymbol{x}_{adv})$ 将根据扰动的变化而变化 $g(\boldsymbol{x})$, 使之偏离原来的解释结果, 即满足 $g(\boldsymbol{x}_{adv}) \neq g(\boldsymbol{x})$.

一种典型的针对解释结果的攻击方法如图 35 所示, 图 35 中 3 个 CNN 表示同一个待解释的预训练 CNN. 其中, Con 表示原图 $\boldsymbol{x}$ 的分类置信度, Exp 表示对应的解释. 使用均方误差损失作为约束, 使对抗图像的分类结果 $f(\boldsymbol{x}_{adv})$ 逼近原图的分类结果 $f(\boldsymbol{x})$, 而解释结果 $g(\boldsymbol{x}_{adv})$ 则逼近目标图的解释结果 $g(\boldsymbol{x}_{target})$, 最终的目标函数是两者的加权和:

$$loss = \lambda_1||g(\boldsymbol{x}_{adv}) - g(\boldsymbol{x}_{target})||^2 + \lambda_2||f(\boldsymbol{x}_{adv}) - f(\boldsymbol{x})||^2 \tag{39}$$

式中, $\boldsymbol{x}_{target}$ 表示用于诱导解释结果的目标图像, λ_1 和 λ_2 为 2 部分的权重参数.

攻击结果如图 36 所示, 图 36(a) 为目标图像 $\boldsymbol{x}_{target}$, 图 36(b) 为原图 $\boldsymbol{x}$, 图 36(c) 为对抗图像 $\boldsymbol{x}_{adv}$, 图 36(e) ~ (g) 分别表示对应的显著图. 由图 36 可以看出, $g(\boldsymbol{x}_{adv})$ 被诱导偏向 $g(\boldsymbol{x}_{target})$, 显示出一只鸟的轮廓. 与此同时, $f(\boldsymbol{x}_{adv})$ 却基本保持不变.

对于使用随机初始化的原图生成的对抗图像图 36(d), 同样可以使用上述攻击方法, 使其对应的显著图 36(h) 被诱导偏向目标图的解释图 36(e), 尽管原图和对抗图像本身没有任务的语义信息. 最终, 分类器对对抗图像图 36(d) 的分类结果图 36(b) 相近, 解释结果与图 36(e) 相近, 但对抗图像图 36(d) 从视觉上看仅是一幅噪声图像. 可见, 显著图解释方法的抗欺骗能力的确存在漏洞, 而目前对于造成这一问题的原因分析仍在探索之中[90].

上述分析显示, 在输入图像未被显著改变、分类结果也保持不变的情形下, 针对分类结果的解释却可以被明显改变而偏向任意目标的解释, 表明可

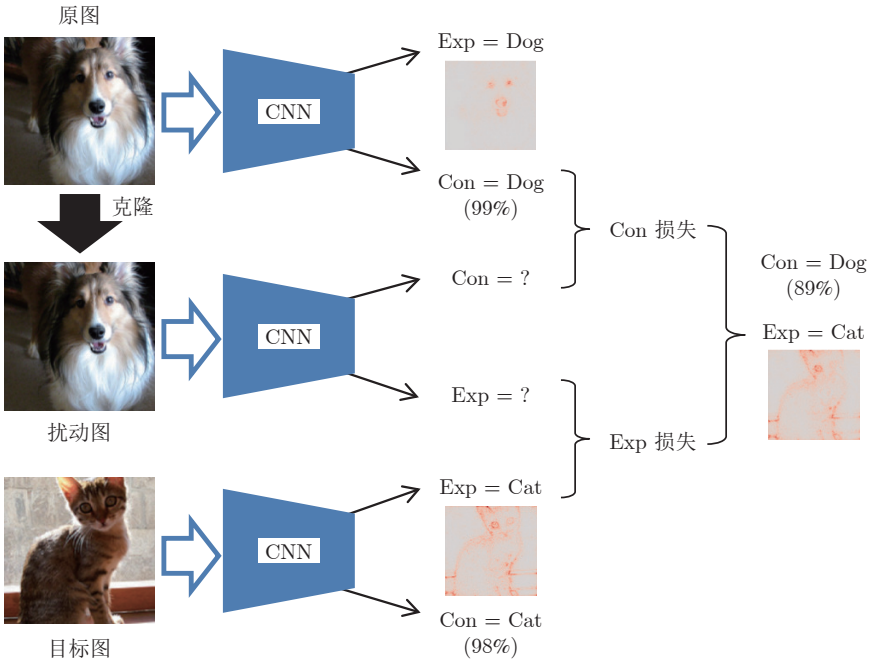


图 35　针对可视化结果的攻击

Fig. 35　Attacks on the visualization results

视化方法存在被欺骗的可能. 文献 [91] 从另一种思路出发, 通过重新微调模型参数, 使微调后的模型的预测结果大致不变, 但解释结果却可以被任意引导. 文献 [92] 对自解释模型所提供的解释的鲁棒性进行了评估, 发现通过创建这样一些对抗性输入, 会使自解释模型提供错误的解释, 表明现有的自解释模型提供的解释鲁棒性并不好, 无法经受住对抗性攻击. 另一方面, 为了提升可视化方法的鲁棒性, 使其不易被误导, 文献 [93] 将显著图应用到模型训练中, 对训练集进行数据增强, 从而训练出归因鲁棒性较好的模型.

4　可视化的应用

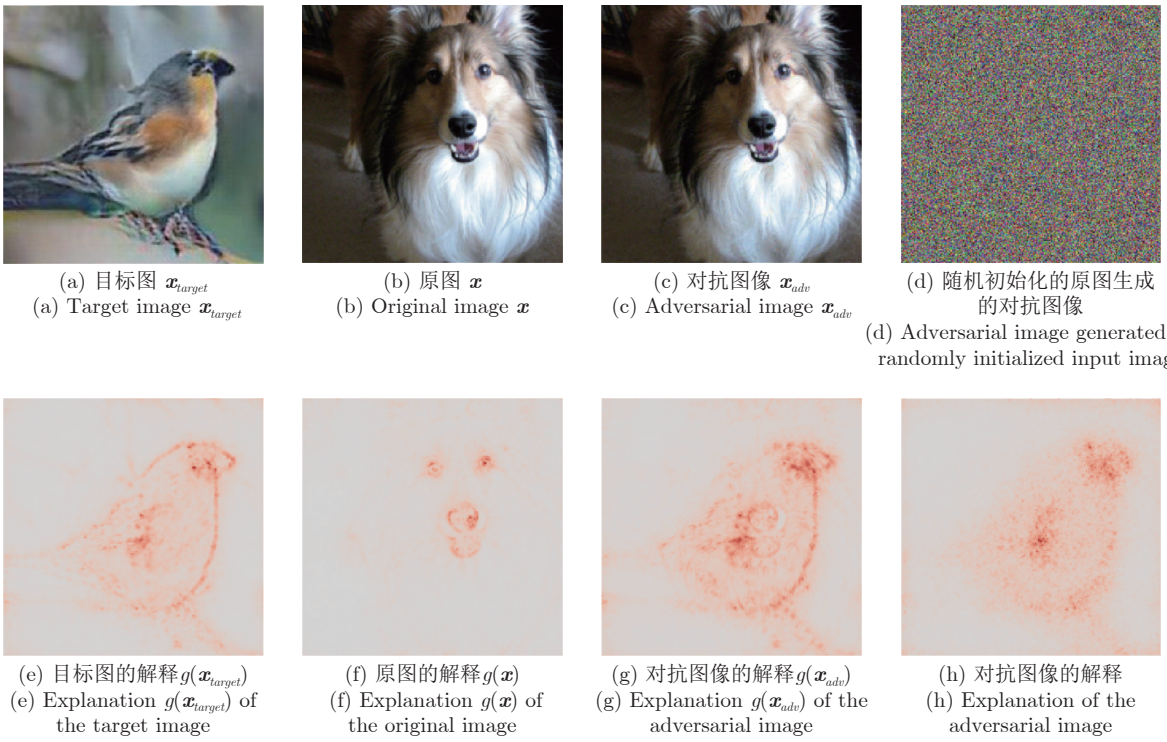
4.1　理解与解释模型

表征可视化是理解 CNN 模型的一种重要途径, 在图像领域应用广泛, 常见于图像分类、场景识别等任务的可视化解释. 本文第 3 节所述的表征可视化方法常用于对基于 CNN 的图像分类器的解释, 例如, AM 方法用于可视化网络对输入图像的偏好, 从另一种角度揭示了网络对何种输入模式的依赖性较强. 注意力掩码能够告诉设计者网络的关注点, 这使其自身具有一定的可解释特性, 因此, 基于注意力掩码的可视化方法不仅可以验证注意力机制自身的有效性, 也常用于观察网络的训练效果.

此外, 表征可视化方法也可以应用在其他类型的数据, 例如, CAM 这类方法具有较好的类别区分性, 能够用来确定与特定输出类别相关联的图像区域, 可在视觉问答模型中帮助定位与问题最相关的图像区域. LRP 方法在制定反向传播规则时依靠网络的权重与激活值, 而非特征图和通道等图像领域的概念. 因此, 它不仅适应于图像识别任务的解释, 还可以用于可视化机器翻译、语音识别[94] 等任务中, 为这些领域的研究者提供了另一种理解模型的途径.

4.2　诊断与优化网络

在 CNN 学习效果诊断和结构优化上, 基于反卷积的可视化能够观察任意层的神经元的激活, 从而分析 CNN 的学习率、卷积核尺寸及步长等重要参数的设计是否达到最优. 文献 [13] 使用基于反卷积的可视化方法对 AlexNet 内部激活进行分析与改进, 进而提出了 ZFNet, 获得了 2013 年 ImageNet 数据集图像分类任务冠军. 这种基于表征可视化的针对性分析和诊断方式, 很大程度上避免了盲目的参数调优. 文献 [95] 利用基于梯度的可视化方法指导单像素的对抗性扰动和对抗性分析, 帮助模型进行对抗性学习. 文献 [88] 则使用显著性方法检测对抗样本, 避免模型受到对抗攻击. 文献 [72] 使用 Grad-CAM 产生的类激活图来观察网络中间层表征, 分析对比不同结构设计对模型训练效果的影响. 此外, CAM 这类方法还可用于提供自注意力, 优化 CNN 的结构设计. 例如, 文献 [73] 和文献 [77]



(a) 目标图 $\boldsymbol{x}_{target}$
(a) Target image $\boldsymbol{x}_{target}$

(b) 原图 $\boldsymbol{x}$
(b) Original image $\boldsymbol{x}$

(c) 对抗图像 $\boldsymbol{x}_{adv}$
(c) Adversarial image $\boldsymbol{x}_{adv}$

(d) 随机初始化的原图生成的对抗图像
(d) Adversarial image generated by randomly initialized input image

(e) 目标图的解释 $g(\boldsymbol{x}_{target})$
(e) Explanation $g(\boldsymbol{x}_{target})$ of the target image

(f) 原图的解释 $g(\boldsymbol{x})$
(f) Explanation $g(\boldsymbol{x})$ of the original image

(g) 对抗图像的解释 $g(\boldsymbol{x}_{adv})$
(g) Explanation $g(\boldsymbol{x}_{adv})$ of the adversarial image

(h) 对抗图像的解释
(h) Explanation of the adversarial image

图 36 使用 GAN 生成的目标图像诱导对 LRP 显著图的攻击[82, 90]

Fig. 36 Using the target image generated by GAN to induce an attack on the LRP saliency map[82, 90]

使用 Grad-CAM 生成自注意力的掩码作为图像蒙版, 用于去除图像中的非重要区域, 并将处理后的图像应用于下阶段的模型训练和推理. 文献 [96] 将 CAM 方法集成到图像转换模型的自注意力模块中, 引导模型关注源域与目标域之间的判别性区域, 从而提升图像转换模型对细节的关注能力.

4.3 其他方面

除了对 CNN 本身的理解与诊断, 可视化方法在其他任务上也有不断拓展与延伸, 例如 CAM 和 Grad-CAM 方法在弱监督目标定位任务上取得了非常好的效果. 文献 [93] 进一步探索了将显著性归因方法产生的显著图作为先验, 应用于弱监督的分割任务上. 在应用领域方面, 可视化方法能够提升对推荐系统决策结果的理解[97], 以及与知识图谱的结合来实现可解释的推荐算法[98]. 对于自动驾驶[99−100]以及智能医疗[101] 等领域, 由于这些领域对于决策风险的承受能力较低, 可视化方法对这些领域应用的现实落地至关重要.

5 存在的难点及发展趋势

5.1 难点分析与趋势展望

近年来, CNN 表征可视化相关研究越来越多, 研究者们提出了各种可视化方法, 极大推动了该领域的进展, 但仍存在一些难点问题有待解决, 本节对其进行了归纳, 并分析了未来可能的研究趋势.

1) 对于可视化方法, 仍存在噪声、稳定性、解释能力有限等问题.

通过对多种可视化方法的实验比较发现, 多数可视化方法生成的热力图含有一定的噪声, 噪声产生的原因仍没有权威统一的解释. 同时, 面对不同图像时的可视化效果不尽相同, 有些图像可能直接导致可视化方法的失效, 而失效的原因尚不清楚, 仍有待进一步的探究. 此外, 面对复杂背景条件的图像、多目标场景、小目标图像等, 受限于模型本身在面对这些情形时的性能约束, 可视化方法的解释效果并不一定好. 未来可能的研究趋势是将可视化方法与其他解释方法的结合, 从不同侧面不同角度解释模型, 从而缓解单一可视化方法解释效果受限的问题.

2) 对于可视化效果的评估, 仍欠缺标准统一的评估方法.

目前很难找到适用于大多数可视化方法的评估标准, 原因在于许多方法的目标并不相同, 也即每种方法对 “可解释性”的理解并不相同, 导致各种可视化方法的解释结果差别较大. 同时, 很多可视化方法自身同样缺乏清晰明确的数学与逻辑机理, 导

致结果难以量化比较. 如果可以从 “可解释性”的概念出发, 统一数个可解释性的标准, 那么对于可视化结果的评估也就有了依据. 同时, 还可以根据可视化方法产生的热力图的特点进行分类评价, 每类热力图使用与之适应的评价标准, 提升其侧重解释某方面的能力.

3) 对于可视化的对象, 细粒度的识别模型难以可视化解释.

可视化方法多应用于对图像分类、目标定位及场景识别等任务的解释, 能够实现对多目标图像中语义级目标的区分. 例如, “Cat”和 “Dog”虽然同属动物, 但是在语义级上属于明显不同的两种动物. 而单独对于 “Cat”这一动物, 实现的不同品种猫的细粒度图像分类, 受限于分类网络自身准确性, 可视化方法很难找到用于区分目标的细节特征, 此时的解释效果非常有限, 甚至对于不同的目标可视化效果始终相同. 与人们的视觉观察及解释能力相差较远. 这一问题或许可以通过视觉解释与语言解释相结合的途径来改善解释效果. 对可视化解释难以描述的细微之处, 辅助加以自然语言描述形式的解释 (比如对猫的颜色、猫耳形状的描述), 能够实现更好的解释效果.

4) 对于可视化解释的完备性, 现有研究中的解释结果与预测结果无法相互印证.

理论上看, 一个完备可靠的解释可以使用户从中推理并得到被解释的预测结果, 而目前的可视化方法仍不具备这一能力, 仅能从预测结果中得到解释结果, 而无法根据解释来推断出模型的预测, 即两者之间的相互印证关系没有被建立起来. 例如, 如果可视化方法给出了错误的解释, 但这一解释恰好符合用户根据预测结果推测的预期解释, 进而使得用户相信了解释的可靠性, 这将对其形成误导. 此时, 若能根据解释结果推断预测结果, 发现推断出的预测结果和实际预测结果不相符合, 则可通过进一步分析发现其中存在的问题, 从而提升用户对可视化方法的信任.

5.2 学界近年来的关注

近年来, 众多人工智能领域顶级会议关注人工智能和深度学习可解释问题, 其中许多涉及到表征可视化方面的前沿研究, 如[102]:

1) IJCAI 2020 Tutorial on Trustworthiness of Interpretable Machine Learning;

2) CVPR 2020 Tutorial on Interpretable Machine Learning for Computer Vision;

3) ICCV 2019 Workshop on Interpretating and Explaining Visual Artificial Intelligence Models;

4) ICLR 2019 Workshop on Safe Machine Learning;

5) CVPR 2019 Workshop on Explainable AI;

6) AAAI 2019 Workshop on Network Interpretability for Deep Learning;

7) IJCAI 2018/2017 Workshop on Explainable Artificial Intelligence;

8) ICML 2018 Workshop on Human Interpretability in Machine Learning;

9) NIPS 2017 Interpretable Machine Learning Symposium.

表 4 列举了可解释性深度学习研究领域的部分综述文献, 对各文献的内容侧重作了简要介绍, 其中包含 CNN 表征可视化的相关内容.

表 4 CNN 表征可视化相关的综述文献统计
Table 4 Review literature statistics related to CNN representation visualization

文献	发表年份	侧重内容
[103]	2016	几种典型的特征可视化方法 (如扰动、反向传播、激活最大化等), 以及相互之间的关系分析
[104]	2017	特征可视化的必要性, 基于反向传播的可视化方法
[105]	2017	模型可视化, 不限于 CNN 可解释性领域
[19]	2018	基于反向传播的可视化方法 (AM、VBP、DTD 和 LRP 等)
[106]	2018	自解释的 CNN
[20]	2018	可解释性的概念, 相关文献分类
[107]	2018	人工智能的可解释性
[102]	2019	机器学习的可解释性方法与评估
[108]	2020	机器学习的可解释性
[109]	2020	深度学习的可解释性
[110]	2020	人工智能的可解释性
[111]	2020	人工智能的可解释性

5.3 开源工具

CNN 可视化的相关开源工具, 一些研究人员在 GitHub 等网站开源了多种方法综合的代码包, 这对于表征可视化研究及迁移到其他任务使用具有重要价值.

文献 [103] 对 2016 年以前的可视化方法作了详细调研和分类整理, 将其中主流方法分为修改输入的方法 (如基于扰动的方法)、反卷积类方法和重建输入的方法 (如激活最大化方法) 三类. 根据这些方法开发了基于 MatConvNet 框架[112] 的 CNN 可视化工具包 FeatureVis, 适用于 Matlab 平台上的 CNN 可视化.

Ozbulak[83] 发布了一个内容丰富的开源代码包, 实现了 10 余种可视化方法, 包括梯度方法 (如 VBP、GAP、Smooth gradient、Integrated gradient 等) 和类激活映射方法 (如 Grad-CAM、Score-CAM 等). 该源码包基于 PyTorch 框架, 已经被许多研究人员关注和使用, 受到领域内好评, 目前仍在更新与拓展中.

韩国科学技术院的 Kim[113] 发布了基于 Tensorflow 框架的可视化源码包, 该源码包含有梯度类方法、CAM 类方法、激活最大化方法等, 配有详细的使用教程, 对各种方法的原理及实现过程的介绍细致, 适合初学者使用.

此外, 佐治亚理工学院的 Wang 等[114] 实现了对 CNN 网络的交互式可视化, 可对 CNN 网络各层的卷积、激活和池化操作的数据流向及中间层特征图进行实时展示, 支持交互式的选择输入图像, 实时观察各层的数据流向及表征情况. 虽然该工具更多关注于 CNN 网络中数据流的走向, 而非解释 CNN 中间层特征的语义, 但也非常有利于理解 CNN 的内部表征.

6 结束语

本文围绕 CNN 表征可视化研究, 详细梳理了该领域近年来相关的文献, 从基础概念及内容、常见方法的分类与比较、效果的评估及应用等方面进行了详细介绍. 其中, 对常见的可视化方法的分类和介绍是本文的重点内容, 该部分详细分析了各种算法的过程, 归纳了每一类方法的特点, 并对它们的效果进行了比较. 最后, 对该领域仍存在的难点和未来的研究趋势作了总结和展望.

随着表征可视化研究的深入, 人们对 CNN 的特征学习和预测机制的理解也会更加深刻. 同时, 其他类型的可解释性方法也在不断发展中, 在它们的共同作用下, 不断推动可解释性深度学习的发展. 期待未来实现可理解的、透明的和高效的深度学习方法.

References

1 Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: 2012. 1106–1114

2 Deng J, Dong W, Socher R, Li L, Li K, L F. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami Beach, Florida, USA: 2009. 248–255

3 Lin T, Maire M, Belongie S J, Hays J, Perona P, Ramanan D, Dollar P, Zitnick C L. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: 2014. 740–755

4 Li M, Zhang T, Chen Y, Smola A J. Efficient mini-batch training for stochastic optimization. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: 2014. 661–670

5 Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint, 2012, arXiv: 1207.0580

6 Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: 2010.807–814

7 Liu Ying, Lei Yan-Bo, Fan Jiu-Lun, Wang Fu-Ping, Gong Yan-Chao, Tian Qi. Survey on image classification technology based on small sample learning. *Acta Automatica Sinica*, 2021, **47**(2): 297−315
(刘颖, 雷研博, 范九伦, 王富平, 公衍超, 田奇. 基于小样本学习的图像分类技术综述. 自动化学报, 2021, **47**(2): 297−315)

8 Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, The MIT Press, 2016.

9 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of The IEEE*, 1998, **86**(11): 2278−2324

10 Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA: 2013.

11 He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: 2015. 1026–103

12 Lin Jing-Dong, Wu Xin-Yi, Chai Yi, Yin Hong-Peng. Structure optimization of convolutional neural networks: A survey. *Acta Automatica Sinica*, 2020, **46**(1): 24−37
(林景栋, 吴欣怡, 柴毅, 尹宏鹏. 卷积神经网络结构优化综述. 自动化学报, 2020, **46**(1): 24−37)

13 Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: 2014. 818–833

14 Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 1–9

15 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014, arXiv: 1409.1556v6

16 He K, Zhang X, Ren S, and Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 770–778

17 Huang G, Liu Z, Maaten L, Weinberger K Q. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: 2017. 2261–2269

18 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 2018. 7132–7141

19 Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2018, **73**: 1−15

20 Gilpin L H, Bau D, Yuan B Z, Bajwa A, Specter M, Kagal L. Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv preprint, 2018, arXiv: 1806.00069

21 Mitros J, Namee B M. A Categorisation of post-hoc explanations for predictive models. arXiv preprint, 2019, arxiv: 1904.02495

22 Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, Müller K R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 2010, **11**(61): 1803−1831

23 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint, 2013, arXiv: 1312.6034

24 Li O, Liu H, Chen C, Rudin C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: 2018. 3530–3537

25 Arik S Ö, Pfister T. ProtoAttend: Attention-based prototypical learning. *Journal of Machine Learning Research*, 2020, **21**(210): 1−3

26 Gulshad S, Smeulders A. Explaining with counter visual attributes and examples. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. Dublin, Ireland: 2020: 35–43

27 Hendricks L A, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T. Generating visual explanations. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, Netherlands: 2016. 3–19

28 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 3156–3164

29 Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 2018. 6077–6086

30 Zhang Q, Wu Y N, Zhu S C. Interpretable convolutional neural networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 2018. 8827–8836

31 Wan A, Dunlap L, Ho D, Yin J, Lee S, Jin H, et al. NBDT: Neural-backed decision trees. arXiv preprint, 2020, arxiv: 2004.00221

32 Ming Y, Cao S, Zhang R, Li Z, Chen Y, Song Y, et al. Understanding hidden memories of recurrent neural networks. In: Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). Phoenix, Arizona, USA: 2017. 13–24

33 Strobelt H, Gehrmann S, Pfister H, Rush A M. LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 2018, **24**(1): 667−676

34 Karpathy A, Johnson J, Li F. Visualizing and understanding recurrent networks. arXiv preprint, 2015, arXiv: 1506.02078

35 Arras L, Montavon G, Müller K R, Samek W. Explaining recurrent neural network predictions in sentiment analysis. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Copenhagen, Denmark: 2017. 159–168

36 Ding Y, Liu Y, Luan H, Sun M. Visualizing and understanding neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC, Canada: 2017. 1150–1159

37 Liu Jian-Wei, Xie Hao-Jie, Luo Xiong-Lin. Research progress on application of generative adversarial networks in various fields. *Acta Automatica Sinica*, 2020, **46**(12): 2500−2536
(刘建伟, 谢浩杰, 罗雄麟. 生成对抗网络在各领域应用研究进展. 自动化学报, 2020, **46**(12): 2500−2536)

38 Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: 2016. 2180–2188

39 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint, 2015, arXiv: 1511.06434

40 Zhu J Y, Krähenbühl P, Shechtman E, Efros A A. Generative visual manipulation on the natural image manifold. In: Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: 2016. 597–613

41 Shen Y, Gu J, Tang X, Zhou B. Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Event: 2020. 9243–9252

42 Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Object detectors emerge in deep scene CNNs. arxiv: 1412.6856, 2014.

43 Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint, 2018, arxiv: 1806.07421

44 Fong R C, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: 2017. 3449–3457

45 Agarwal C, Schonfeld D, Nguyen A. Removing input features via a generative model to explain their attributions to classifier′s decisions. arXiv: 1910.04256, 2019.

46 Chang C H, Creager E, Goldenberg A, Duvenaud D. Explaining image classifiers by counterfactual generation. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: 2019.

47 Fong R, Patrick M, Vedaldi A. Understanding deep networks via extremal perturbations and smooth masks. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea: 2019. 2950–2958

48 Wagner J, Kohler J M, Gindele T, Hetzel L, Wiedemer J T, Behnke S. Interpretable and fine-grained visual explanations for convolutional neural networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: 2019. 9097–9107

49 Vedaldi A. Understanding models via visualizations and attribution [Online], available: https://interpretablevision.github.io/slide/iccv19_vedaldi_slide.pdf, 2019.

50 Springenberg J T, Dosovitskiy A, Brox T, Riedmiller M A. Striving for simplicity: The all convolutional net. arXiv preprint, 2014, arXiv: 1412.6806

51 Sundararajan M, Taly A, Yan Q. Gradients of counterfactuals.

arXiv: 1611.02639, 2016.

52 Smilkov D, Thorat N, Kim B, Viegas F B, Wattenberg M. Smoothgrad: Removing noise by adding noise. arXiv preprint, 2017, arXiv: 1706.03825

53 Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, NSW, Australia: 2017. 3319–3328

54 Kim B, Seo J, Jeon S, Koo J, Choe J, Jeon T. Why are Saliency maps noisy? Cause of and solution to noisy saliency maps. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea: 2019. 4149–4157

55 Hooker S, Erhan D, Kindermans P J, Kim B. A benchmark for interpretability methods in deep neural networks. In: Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, Canada: 2019. 9737–9748

56 Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In: Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada: 2018.

57 Rieger L, Hansen L K. Aggregating explanation methods for stable and robust explainability. arXiv: 1903.00519, 2019.

58 Bach S, Binder A, Montavon G, Klauschen F, Müller K, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One*, 2015, **10**(7): 0130140

59 Gu J, Yang Y, Tresp V. Understanding individual decisions of cnns via contrastive backpropagation. In: Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: 2018. 119–134

60 Iwana B K, Kuroki R, Uchida S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea: 2019. 4176–4185

61 Montavon G, Lapuschkin S, Binder A, Samek W, Müller K R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 2017, **65**: 211−222

62 Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 2921–2929

63 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: 2017. 618–626

64 Selvaraju R R, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? arXiv preprint, 2016, arXiv: 1611.07450

65 Chattopadhay A, Sarkar A, Howlader P, Balasubramanian V N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, Nevada, USA: 2018. 839–847

66 Wang H, Du M, Yang F, Zhang Z. Score-CAM: Improved visual explanations via score-weighted class activation mapping. arXiv preprint, 2019, arXiv: 1910.01279

67 Patro B, Lunayach M, Patel S, Namboodiri V. U-CAM: visual explanation using uncertainty based class activation maps. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea: 2019. 7444–7453

68 Omeiza D, Speakman S, Cintas C, Weldermariam K. Smooth Grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint, 2019, arXiv: 1908.01224

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

70 Mahendran A, Vedaldi A. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 2016, **120**(3): 233−255

71 Yosinski J, Clune J, Nguyen A M, Fuchs T J, Lipson H. Understanding neural networks through deep visualization. arXiv preprint, 2015, arXiv: 1506.06579

72 Woo S, Park J, Lee J Y, Kweon I S. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: 2018. 3–19

73 Li K, Wu Z, Peng K C, Ernst J, Fu Y. Tell me where to look: Guided attention inference network. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 2018. 9215–9223

74 Fukui H, Hirakawa T, Yamashita T, Fujiyoshi H. Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: 2019. 10705–10714

75 Zhang J, Lin Z L, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. In: Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 543–559

76 Hua Y, Mou L, Zhu X X. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2019, **149**: 188−199

77 Li J, Lin D, Wang Y, Xu G, Ding C. Deep discriminative Representation learning with attention map for scene classification. arXiv preprint, 2019, arXiv: 1902.07967

78 Ribeiro M T, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. San Diego, USA: 2016. 97–101

79 Lundberg S M, Lee S I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: 2017. 4768–4777

80 Shapley L S. *A Value for N-Person Games. Contributions to The Theory of Games (AM-28)*. Princeton: Princeton University Press, 1953. 2: 307–317

81 Erhan D, Bengio Y, Courville A, Vincent P. Visualizing Higher-layer Features of a Deep Network, Technical Report 1341, Department of Computer Science and Operations Research, University of Montreal, Canada, 2009

82 Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Proceedings of the Annual Conference on Neural Information Processing Systems. Barcelona, Spain: 2016. 3387–3395

83 Ozbulak U. PyTorch CNN Visualizations [Online], available: https://github.com/utkuozbulak/pytorch-cnn-visualizations, 2019.

84 Zhang Q, Wang W, Zhu S C. Examining CNN representations with respect to dataset bias. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: 2017. 4464–4473

85 Adebayo J, Gilmer J, Muelly M C, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Proceedings of the Annual Conference on Neural Information Processing Systems. Montréal, Canada: 2018. 9505–9515

86 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: Proceedings of the 2014 ICLR International Conference on Learning Representations. Banff, Canada: 2014.

87 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 2015 ICLR International Conference on Learning Representations. San Diego, USA: 2015.

88 Gu J, Tresp V. Saliency Methods for explaining adversarial attacks. arXiv preprint, 2019, arXiv: 1908.08413

89 Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA: 2019. 3681–3688

90 Dombrowski A K, Alber M, Anders C, Ackermann M, Müller K R, Kessel P. Explanations can be manipulated and geometry is to blame. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: 2019. 13589–13600

91 Heo J, Joo S, Moon T. Fooling neural network interpretations via adversarial model manipulation. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: 2019. 2925–2936

92 Zheng H, Fernandes E, Prakash A. Analyzing the interpretability robustness of self-explaining models. arXiv preprint, 2019, arXiv: 1905

93 Singh M, Kumari N, Mangla P, Sinha A, Balasubramanian V N, Krishnamurthy B. On the benefits of attributional robustness. arXiv preprint, 2019, arXiv: 1911.13073

94 Krug A, Stober S. Introspection for convolutional automatic speech recognition. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: 2018. 187–199

95 Kumar D, Daya I B, Vats K, Feng J, Taylor G W, Wong A. Beyond explainability: Leveraging interpretability for improved adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: 2019. 16–19

96 Kim J, Kim M, Kang H, Lee K H. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: 2020.

97 Tan Y, Zhang M, Liu Y, Ma S. Rating-boosted latent topics: Understanding users and items with ratings and reviews. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: 2016. 2640–2646

98 Zhang Y, Chen X. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, 2020, **14**(1): 1–101

99 Bojarski M, Choromanska A, Choromanski K, Firner B, Jackel L, Muller U, et al. VisualBackProp: Visualizing CNNs for autonomous driving. arXiv preprint, 2016, arxiv: 1611.05418

100 Kim J, Canny J. Interpretable learning for self-driving cars by visualizing causal attention. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: 2017. 2961–2969

101 Zhang Z, Xie Y, Xing F, McGough M, Yang L. MDNet: A semantically and visually interpretable medical image diagnosis network. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: 2017. 3549–3557

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

103 Grün F, Rupprecht C, Navab N, Tombari F. A taxonomy and library for visualizing learned features in convolutional neural networks. arXiv preprint, 2016, arXiv: 1606.07757

104 Samek W, Wiegand T, Müller K R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint, 2017, arXiv: 1708.08296

105 Seifert C, Aamir A, Balagopalan A, Jain D, Sharma A, Grottel S, et al. Visualizations of deep neural networks in computer vision: A survey. *Studies in Big Data*, 2017: 123–144

106 Zhang Q, Zhu S. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, 2018, **19**(1): 27–39

107 Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 2018, **6**: 52138–52160

108 Samek W, Montavon G, Lapuschkin S, Anders C J, Müller K. Toward interpretable machine learning: Transparent deep neural networks and beyond. arXiv preprint, 2020, arXiv: 2003.07631

109 Xie N, Ras G, Gerven M van, Doran D. Explainable deep learning: A field guide for the uninitiated. arXiv preprint, 2020, arXiv: 2004.14545

110 Arrieta A B, Díaz-Rodríguez N, Ser J D, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020, **58**: 82–115

111 Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint, 2020, arXiv: 2006.11371

112 Vedaldi A, Lux M, Bertini M. MatConvNet: CNNs are also for Matlab users. *ACM Sigmultimedia Records*, 2018, **10**(1): 9

113 Kim B. Understanding NN [Online], available: https://github.com/1202kbs/Understanding-NN, August 22, 2020.

114 Wang Z J, Turko R, Shaikh O, Park H, Das N, Hohman F, et al. CNN Explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2021, **27**: 1396–1406

司念文 信息工程大学信息系统工程学院博士研究生. 主要研究方向为深度学习的安全性与可解释性.
E-mail: snw1608@163.com
(**SI Nian-Wen** Ph.D. candidate at the College of Information System Engineering, Information Engineering University. His research interest covers deep learning security and interpret ability.)

张文林 信息工程大学信息系统工程学院副教授. 主要研究方向为深度学习和语音识别. 本文通信作者.
E-mail: zwlin_2004@163.com
(**ZHANG Wen-Lin** Associate professor at the College of Information System Engineering, Information Engineering University. His research interest covers deep learning and speech recognition. Corresponding author of this paper.)

屈 丹 信息工程大学信息系统工程学院教授. 主要研究方向为机器学习, 深度学习和语音识别.
E-mail: qudanqudan@163.com
(**QU Dan** Professor at the College of Information System Engineering, Information Engineering University. Her research interest covers machine learning, deep learning and speech recognition.)

罗向阳 信息工程大学网络空间安全学院教授. 主要研究方向为人工智能与信息安全.
E-mail: xiangyangluo@126.com
(**LUO Xiang-Yang** Professor at the College of Cyberspace Security, Information Engineering University. His research interest covers artificial intelligence and information security.)

常禾雨 信息工程大学密码工程学院博士研究生. 主要研究方向为深度学习与行人重识别.
E-mail: okaychy@163.com
(**CHANG He-Yu** Ph.D. candidate at the College of Cryptographic Engineering, Information Engineering University. Her research interest covers deep learning and person re-identification.)

牛 铜 信息工程大学信息系统工程学院副教授. 主要研究方向为深度学习和语音识别.
E-mail: jerry_newton@sina.com
(**NIU Tong** Associate professor at the College of Information System Engineering, Information Engineering University. His research interest covers deep learning and speech recognition.)