基于逐层增量分解的深度网络神经元相关性解释方法

陈艺元 [1, 2] 李建威 [1, 2] 邵文泽 [1, 2] 孙玉宝 [3]

摘 要 神经网络的黑箱特性严重阻碍了人们关于网络决策的直观分析与理解. 尽管文献报道了多种基于神经元贡献度分配的决策解释方法, 但是现有方法的解释一致性难以保证, 鲁棒性更是有待改进. 本文从神经元相关性概念入手, 提出一种基于逐层增量分解的神经网络解释新方法 LID-Taylor (Layer-wise increment decomposition), 且在此基础上先后引入针对顶层神经元相关性的对比提升策略, 以及针对所有层神经元相关性的非线性提升策略, 最后利用交叉组合策略得到最终方法 SIG-LID-IG, 实现了决策归因性能的鲁棒跃升. 通过热力图对现有工作与提出方法的决策归因性能做了定性定量评估. 结果显示, SIG-LID-IG 在神经元的正、负相关性的决策归因合理性上均可媲美甚至优于现有工作. SIG-LID-IG 在多尺度热力图下同样取得了精确性更高、鲁棒性更强的决策归因.

关键词 神经网络, 可解释性, 决策相关性, 逐层相关性传播, 类激活图, 积分梯度

引用格式 陈艺元, 李建威, 邵文泽, 孙玉宝. 基于逐层增量分解的深度网络神经元相关性解释方法. 自动化学报, 2024, **50**(10): 2049−2062

DOI 10.16383/j.aas.c230651 **CSTR** 32138.14.j.aas.c230651

Layer-wise Increment Decomposition-based Neuron Relevance Explanation for Deep Networks

CHEN Yi-Yuan[1, 2] LI Jian-Wei[1, 2] SHAO Wen-Ze[1, 2] SUN Yu-Bao[3]

Abstract The black box nature of deep neural networks seriously hinders one′s intuitive analysis and understanding of network decision-making. Although various decision explanation methods based on neural contribution allocation have been reported in the literature, the consistency of existing methods is difficult to ensure, and their robustness still needs improvement. This article starts with the concept of neuron relevance and proposes a new neural network explanation method LID-Taylor (layer-wise increment decomposition). Aiming at LID-Taylor, a contrast lifting strategy for top-layer neuron relevance and a non-linear lifting strategy for all-layer neuron relevance are introduced, respectively. Finally, a cross combination strategy is applied, obtaining the final method SIG-LID-IG and achieving a robust leap in decision attribution performance. Both qualitative and quantitative evaluation have been conducted via heatmaps on the decision attribution performance of existing works and the proposed method. Results show that SIG-LID-IG is comparable or even superior to existing works in the rationality of positive and negative relevance of neurons in decision-making attribution. SIG-LID-IG has also achieved better accuracy and stronger robustness in decision-making attribution in terms of multi-scale heatmaps.

Key words Neural network, explainability, decision relevance, layer-wise relevance propagation (LRP), class activation map, integrated gradients (IG)

Citation Chen Yi-Yuan, Li Jian-Wei, Shao Wen-Ze, Sun Yu-Bao. Layer-wise increment decomposition-based neuron relevance explanation for deep networks. *Acta Automatica Sinica*, 2024, **50**(10): 2049−2062

收稿日期 2023-10-23 录用日期 2024-04-29
Manuscript received October 23, 2023; accepted April 29, 2024
国家自然科学基金 (61771250, 61972213, 62276139, U2001211), 青蓝工程资助
Supported by National Natural Science Foundation of China (61771250, 61972213, 62276139, U2001211) and Qing Lan Project
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 智能信息处理与通信技术省高校重点实验室 (南京邮电大学) 南京 210003 2. 南京邮电大学通信与信息工程学院 南京 210003 3. 南京信息工程大学教育部数字取证工程研究中心 南京 210044
1. Jiangsu Key Laboratory of Intelligent Information Processing and Communication Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003 2. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003 3. Engineering Research Center for Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044

当下, 以卷积神经网络 (Convolutional neural networks, CNN)[1−4] 为代表的深度学习方法在图像分类[1]、目标识别[5−6] 等领域取得了巨大的成功. 然而, 神经网络具有明显的黑箱特性, 内部神经元的复杂组合严重阻碍了人们关于网络决策的直观分析与理解[7]. 为此, 如何打开黑箱, 可信地再现决策过程, 成为近年来众多神经网络解释方法的要旨所在.

可解释性研究的目的之一是使人类直观地理解神经网络提取的特征并作出决策的过程. 其中, 一类代表性解释方法注重于直观感受提取的特征[8], 以便于理解内部神经元学习的深度语义表示. 例如

激活最大化方法 (Activation maximization, AM)[9] 逆转中间层神经元的特征提取过程, 将内部神经元映射至输入域. AM 通过优化搜索使神经元激活值最大化的像素域图像, 将特征表示可视化为潜在的输入图像.

另外一类工作注重于特征决策, 以特征贡献度分配的方式定性或定量刻画网络的决策过程. 如, 以类激活图 (Class activation mapping, CAM)[10] 为代表的系列方法[11−13] 利用中间特征的热力图进行决策可视化, 以直观理解决策结果与决策区域之间的语义相关性. 而以逐层相关性传播 (Layer-wise relevance propagation, LRP)[14] 为代表的贡献度分配方法[15−20], 为每个神经元计算关于网络决策的相关性得分, 可以实现复杂特征决策过程的定量化解释. 简单而言, LRP 方法为所有神经元显式分配贡献度, 而决策结果是神经元贡献度之和. LRP 的神经元相关性也可用于生成热力图, 可视化神经元的贡献. 积分梯度 (Integrated gradients, IG)[21] 与 SHAP (Shapley additive explanations)[22] 等方法进一步构建了决策归因的公理化系统, 同时探讨了贡献度得分应满足的若干必要性质.

本文注意到, 尽管文献报道了多种基于贡献度分配的网络决策解释方法, 但是不同方法的解释一致性难以保证, 精确性和鲁棒性更是有待改进, 打开“黑箱”依然任重道远. 精确性指的是神经元相关性解释方法即神经元相关性 (贡献度) 归因的准确度, 而鲁棒性则是神经元相关性解释方法在噪声、平移抖动与缩放等变换下的抗干扰性.

本文的研究动机源于 LRP 方法, 将从神经元的相关性概念入手, 通过对神经元的贡献度得分的本原探索, 提出一种基于逐层增量分解的神经网络解释新方法 (Layer-wise increment decomposition, LID), 且在此基础上进一步探讨了神经元的顶层相关性与中间层相关性计算方法的提升策略与交叉组合策略. 具体而言, 本文工作归纳如下:

1) 首先, 提出一种基于泰勒的逐层增量分解新方法 LID-Taylor, 为神经元的相关性解释和决策贡献度分配提供了一种新视角.

2) 其次, 引入一种针对顶层神经元相关性的对比提升策略, 通过更为合理的正、负贡献度分配, 助力 LID-Taylor 实现决策归因的类区分性.

3) 之后, 引入一种针对所有层神经元相关性的非线性提升策略, 实现基于梯度积分的增量计算, 助力 LID-Taylor 逐层改善决策归因的精确性.

4) 最后, 针对顶层神经元相关性采取对比与非线性交叉组合提升策略, 助力 LID-Taylor 决策归因性能的鲁棒跃升, 本文将最终的方法称为 SIG-LID-IG.

在实验验证部分, 本文通过热力图对现有工作与本文方法的决策归因性能做了定性与定量评估. 第 4.2 节与第 4.3 节的实验结果表明, 本文 SIG-LID-IG 在神经元的正相关性、负相关性的决策归因合理性上均可媲美甚至优于现有方法, 具有更高的精确性. 第 4.4 节的实验表明, 本文方法在抗扰动方面还具有很好的鲁棒性. 此外, 相较于 Layer-CAM[12], SIG-LID-IG 在多尺度热力图下同样取得了精确性更高、鲁棒性更强的决策归因.

1 相关工作

以下, 对 CNN 模型在分类任务下常见函数符号表示作统一表述. 输入图像记为 X, 网络层号记为 $0,\ 1,\ \cdots,\ L,\ L+1$. 定义输入层为第 0 层, 即 $Y^0 = X$. 前一层输出是后一层输入, Y^l 表示第 l 层全体神经元及其输出, Y_j^l 表示第 l 层第 j 个神经元. 将 logits 层 (最后的 FC 层, 全连接层) 的输出 Y^L 记为 Z, 而将 Softmax 层 (logits 层的下一层) 的输出概率记为 P. 对于输入图像 X, 假设对应类别 c 的 logits 层决策函数为 $f(X) = Z_c$, c 可以是 X 的真实标签或者其他任意类别. 在此设定下, 本文着重讨论 Y_j^l 对 $f(X)$ 的贡献度得分或语义相关性得分, 记为 $R(Y_j^l)$.

1.1 LRP

LRP[15] 首次引入了神经元相关性的概念, 通过自下而上、逐层递归的方式, 可计算每个神经元与类别 c 的决策相关性. 特别地, 如式 (1) 所示, CNN 的全体神经元与决策函数 $f(X)$ 的语义相关性应满足逐层守恒性.

$$f(X) = \cdots = \sum_j R(Y_j^l) = \cdots = \sum_k R(Y_k^0) \tag{1}$$

分配规则[15] 是 LRP 针对逐层相关性计算提出的两个重要规则之一. 即, 对于第 l 层的第 j 个神经元的语义相关性得分 $R(Y_j^l)$, 一部分将分配到第 $l-1$ 层第 i 个神经元, 定义为相关性信息 $R_{j\to i}^{l\to l-1}$. 神经元 Y_j^l 分配的相关性信息是守恒的, 如式 (2) 所示.

$$R(Y_j^l) = \sum_i R_{j\to i}^{l\to l-1} \tag{2}$$

吸收规则[15] 是逐层相关性计算的另一规则. 即, 第 $l-1$ 层第 i 个神经元的相关性得分 $R(Y_i^{l-1})$ 是由第 l 层所有神经元传递而来的相关性信息的总和, 如式 (3) 所示.

$$\sum_j R_{j\to i}^{l\to l-1} = R(Y_i^{l-1}) \tag{3}$$

1.2 LRP-0

相关性得分的逐层计算是 LRP 的核心, 而 LRP-0[15] 是 LRP 给出的第一个具体规则, 根据输入神经元对输出神经元的贡献按比例计算相关性信息. 值得注意的是, LRP-0 是 LRP 针对卷积层、全连接层提出的逐层计算规则, 而对于最大池化层 (Maxpool), LRP 提出了如 Winner-take-all (WTA) 等规则[15], 为最大值元素分配所有相关性; 对于激活层则跳过计算. 为下文叙述方便, 式 (4) 给出了全连接层前向传播过程, W_{ji}^l 与 B_j^l 是第 l 层的权重与偏移.

$$Y_j^l = \sum_i W_{ji}^l \cdot Y_i^{l-1} + B_j^l \tag{4}$$

当选择决策函数为 $f(X) = Z_c$, LRP-0 随之进行顶层相关性的初始化, 即, 神经元 Y_c^L 的相关性 $R(Y_c^L)$ 设定为 Z_c 本身, 而其他类别的相关性则为零. 顶层相关性可以统一表示为向量 $R(Y^L) = e_c \odot Z$, 式中 e_c 是单位向量, $\odot$ 是逐元素相乘. LRP-0 中间层的相关性信息 $R_{j\to i}^{l\to l-1}$ 的逐层计算规则如式 (5) 所示, 可由输出神经元 Y_j^l 的相关性得分 $R(Y_j^l)$ 乘以分配比例简单得到. 其中, 分配比例由加权输入 $W_{ji}^l Y_i^{l-1}$ 除以输出 Y_j^l 而得, 显然加权输入越大, 分配比例越大.

$$R_{j\to i}^{l\to l-1} = R(Y_j^l) \cdot \frac{W_{ji}^l \cdot Y_i^{l-1}}{Y_j^l} \tag{5}$$

1.3 DeepLIFT

DeepLIFT (Deep learning important features)[23] 是一种基于参考点 $\hat{X}$ 的贡献度分配方法. 参考点 $\hat{X}$ 是输入 X 的对比, 通常选择为零. 与 LRP 逐层分解决策函数 $f(X)$ 所不同的是, DeepLIFT 是将决策增量 $\Delta f = f(X) - f(\hat{X})$ 逐层分解为神经元的贡献度. 具体而言, DeepLIFT 使用参考点 $\hat{X}$ 前向传播, 神经元 Y_j^l 的参考输出记为 $\hat{Y}_j^l$, 而对应输出增量 $\Delta Y_j^l = Y_j^l - \hat{Y}_j^l$ 的贡献度得分记为 $R(Y_j^l, \Delta Y_j^l)$. 类似 LRP, DeepLIFT 提出如式 (6) 所示的贡献度守恒. 在不混淆的情况下, 贡献度 $R(Y_j^l, \Delta Y_j^l)$ 也简记为 $R(Y_j^l)$.

$$\Delta f = \sum_j R(Y_j^l, \Delta Y_j^l) \tag{6}$$

类似 LRP-0, DeepLIFT 针对卷积层、全连接层提出如式 (7) 所示的相关性信息 $R_{j\to i}^{l\to l-1}$ 计算规则, 其他层的相关性信息则需另行计算. 其中, 分配比例由加权输入增量 $W_{ji}^l \Delta Y_i^{l-1}$ 除以输出增量 ΔY_j^l 而得.

$$R_{j\to i}^{l\to l-1} = R(Y_j^l) \cdot \frac{W_{ji}^l \cdot \Delta Y_i^{l-1}}{\Delta Y_j^l} \tag{7}$$

然而, 不管是 LRP-0 还是 DeepLIFT, 相关性信息计算式中的除数可能为零, 有可能引发数值不稳定性. 为此, LRP 在分母上添加一个微小常数, 引入 LRP-ϵ [15] 规则; DeepLIFT (Reveal Cancel)[23] 则通过正、负贡献度区分的方式 (Separating positive and negative contributions) 实现计算的数值稳定性, 或直接令 $R_{j\to i}^{l\to l-1}$ 为零. 与 LRP-0、DeepLIFT 不同的是, 本文将有效避免涉及式 (5) 或式 (7) 中的除法运算, 转而探寻一种相关性信息计算的新方法.

1.4 DTD

DTD (Deep taylor decomposition)[18] 是一种基于泰勒分解的相关性计算方法. 与基于相关性分配的 LRP 不同, DTD 假设任意一个上层神经元存在一个可表示为以下层神经元为变量的相关性函数 $R(Y_j^l)(Y^{l-1})$, 直接对相关性函数进行递归的泰勒分解. DTD 需要寻找每个神经元的相关性函数的根点, 将相关性函数一阶泰勒展开项递归地传递下去, 如式 (8) 所示, 其中 $\hat{Y}^{l-1}(j)$ 是第 j 个神经元的根点, 使得 $R(Y_j^l)(\hat{Y}^{l-1}(j)) = 0$, 而不同神经元的根点是不同的.

$$R_{j\to i}^{l\to l-1} = \frac{\partial R(Y_j^l)}{\partial Y_i^{l-1}}\Big|_{\hat{Y}^{l-1}(j)} \cdot (Y_i^{l-1} - \hat{Y}_i^{l-1}(j)) \tag{8}$$

DTD 提出了一种适用于 ReLU 网络的 Training-free 相关性模型, 在添加 ReLU 激活函数的 FC 层上递归地使用相关性模型 $R(Y_j^l)(Y_i^{l-1}) = C_j^l Y_j^l = ReLU(C_j^l(\sum_i W_{ji}^l Y_i^{l-1} + B_j^l))$[17] 表示每个隐藏层神经元的相关性函数, 其中 C_j^l 是非负乘项. 这种方法允许网络递归地使用相同的 z^+ 规则, 如式 (9) 所示, 其中 $W_{ji}^{l+} = \max(W_{ji}^l, 0)$. z^+ 规则与 LRP-0 的计算过程非常相似, 又称为 LRP-ZP, 使得 DTD 的计算与 LRP 系列兼容.

$$R(Y_i^{l-1}) = \sum_j R(Y_j^l) \cdot \frac{W_{ji}^{l+} \cdot Y_i^{l-1}}{\sum_k W_{jk}^{l+} \cdot Y_k^{l-1}} \tag{9}$$

DTD 引入的相关性模型理论较为复杂, 模型构建的假设较多, 缺少充分的解释, 具有一定的局限性. 1) DTD 假设决策函数总是非负的, 并且当 $f(X) = 0$ 时, 表示不存在该类目标, 这与现有的神经网络矛盾. 例如 logits 具有偏移不变性, 对所有元素加以任意相同常数 (负数), 预测概率不变. 又例如对于二分类概率 probability, 无法分类的情况

其实是概率为 0.5. 2) DTD 假设神经元的贡献也是非负的. 虽然 ReLU 网络的激活总是正值, 但是神经元有可能做出负贡献. z^+ 规则丢失了负相关性, 同时热力图缺少类区分度[19].

DTD 成功地运用统一的泰勒思想, 通过递归 Training-free 模型实现了统一的 LRP-ZP 规则. 根据这项研究的启发, 本文从泰勒展开的视角, 重新探寻简洁统一的相关性解释方法.

1.5 SG-LRP

对比相关性方法 C-LRP (Contrastive-LRP)[19] 是 LRP 方法的进阶, 指出 LRP 对于不同类别 c 所生成的热力图高度相似. 与 LRP 不同的是, C-LRP 将其他类别的顶层相关性设置为相同的负值, 以消减关于其他类的相关性, 这在一定程度上可实现热力图的类区分性. SG-LRP (Softmax gradient-LRP)[20] 对 C-LRP 做了进一步延伸讨论, 提出如式 (10) 所示的基于 Softmax 梯度的顶层相关性. 其中, 对应非目标类的神经元顶层相关性为 $-P_cP_i$. 需要注意的是, 与 LRP-0 不同, SG-LRP 在中间层沿用了 DTD 的 z^+ 规则[18], 这种方法记为 SG-LRP-ZP. 容易知道, SG-LRP 旨在通过与目标类无关的语义信息的自适应抑制改善热力图的类区分性.

$$R(Y_i^L)=\frac{\partial P_c}{\partial Z_i}=\begin{cases}(1-P_c)\cdot P_i, & i=c\\ -P_c\cdot P_i, & i\neq c\end{cases} \tag{10}$$

然而, 直观上, SG-LRP 存在两个明显的缺陷. 1) 非目标类的顶层相关性均为负值存在归因错误. 不难理解, 与参考点比较, 当非目标类 i 的 logits 得分 Z_i 上升, 目标类的概率 P_c 将下降, 此时非目标类 Z_i 确实产生了负贡献. 但是, 当非目标类的 logits 得分下降, 目标类的概率将上升, 非目标类应产生正贡献而不是负贡献. 2) Softmax 梯度存在饱和问题. 如图 1 所示, 当某维度的输入持续增大, Softmax 函数趋于平缓, 概率相应趋近于 1, Softmax 梯度自然趋于饱和. 因而由式 (10) 知, 不管是目标类还是非目标类, 此时神经元的顶层相关性均将随之消失.

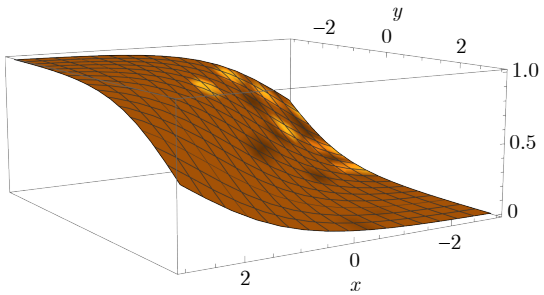


图 1 Softmax 函数

Fig. 1 The Softmax function

1.6 IG

与 DeepLIFT 类似, 积分梯度[21] 是基于参考点的贡献度分配方法. 但与 DeepLIFT 不同的是, IG 最初提出时只是将决策函数 $f(X)$ 归因到输入 $X=(X_1,\ X_2,\ \cdots,\ X_i,\ \cdots)$ 的各个分量, 如式 (11) 所示. 其中, $R(X_i)$ 通过梯度分量 $\frac{\partial f(X)}{\partial X_i}$ 在 $\hat{X}\to X$ 的路径积分计算得到.

$$R(X_i)=\int_{\hat{X}\to X}\frac{\partial f(X)}{\partial X_i}\mathrm{d}X_i \tag{11}$$

要指出的是, 虽然 $\hat{X}\to X$ 有多种路径, 但是 IG 方法遵循简单而有效的原则, 选择了 $\hat{X}$ 和 X 之间的直线路径对梯度进行积分. 此外, 尽管式 (11) 定义的是输入层决策归因, 但是不难理解中间层积分梯度可以类似方式计算.

2 基于泰勒的逐层增量分解

作为一种神经元相关性解释新方法, 基于泰勒的逐层增量分解 LID-Taylor 同样面向决策增量 Δf, 从泰勒展开的视角重新探讨逐层神经元的决策贡献度 $R(Y_j^l,\ \Delta Y_j^l)$, 本文亦称之为增量相关性. LID-Taylor 有两个特点: 1) 研究关于增量的相关性, 源于函数的零点并非输入为零的情况, 例如当函数 $f(0)\neq 0$, 此时净输入 X 不能完全构成净输出 $f(X)$ 的总贡献. 2) 参考了 DTD 泰勒展开的思想, 但放弃了相关性函数, 直接对神经元进行一阶泰勒展开, 继而沿用并推广 LRP 相关性的分配吸收规则. 与 LRP 和 DeepLIFT 等系列方法不同的是, LID-Taylor 实现了所有层神经元的统一相关性计算.

式 (12) 给出了输出增量 ΔY_j^l 关于 Y_i^{l-1} 的泰勒展开, 其中 $D_{ji}^l=\frac{\partial Y_j^l}{\partial Y_i^{l-1}}$ 是 Y_j^l 在 Y_i^{l-1} 处的局部偏导数, ϵ 则是高阶余项.

$$\Delta Y_j^l=Y_j^l-\hat{Y}_j^l=\sum_i D_{ji}^l\cdot\Delta Y_i^{l-1}+\epsilon \tag{12}$$

根据式 (12), 泰勒分解能够以前向传播的方式, 实现任意网络层的增量 ΔY_j^l 的直接计算, 从而对照式 (4), LID-Taylor 可逐层计算包括全连接、卷积、激活、最大池化等任意网络层的增量相关性. 特别地, 对于卷积层和全连接层, 容易证明 $D_{ji}^l=W_{ji}^l$, 所以 D_{ji}^l 与 W_{ji}^l 均刻画的是网络层的线性部分. 而激活与最大池化层等非线性因素, 则是产生高阶余项 ϵ 的直接原因.

2.1 LID-Taylor

根据泰勒分解式 (12), 增量 ΔY_j^l 近似为线性贡

献 $D_{ji}^{l}\Delta Y_i^{l-1}$ 的总和, 即 $\Delta Y_j^l \approx \sum_i D_{ji}^{l}\Delta Y_i^{l-1}$. 自然地, LID-Taylor 的相关性信息可由线性贡献 $D_{ji}^{l}\Delta Y_i^{l-1}$ 占输出增量 ΔY_i^l 的比例计算而来, 如式 (13) 所示. 与前文式 (7) 相比, 此处不同之处仅在于 W_{ji}^l 替换为 D_{ji}^l. 因此, LID-Taylor 是 DeepLIFT 的一般推广.

$$R_{j\to i}^{l\to l-1} = R(Y_j^l)\cdot\frac{D_{ji}^{l}\cdot\Delta Y_i^{l-1}}{\Delta Y_j^l} \tag{13}$$

进一步地, 本文利用式 (14) 的矩阵−向量形式简洁表达了增量相关性的逐层传播规则. 1) 假设第 l 层与第 $l-1$ 层各有 N 和 M 个神经元, $R(Y^{l-1})$ 和 $R(Y^l)$ 是对应层的增量相关性, D^l 是 $N\times M$ 维的局部雅可比矩阵, 而“$\odot$”与“/”代表逐元素乘法和除法, “$\odot$”可省略. 2) 式 (14) 右侧遵循了 Einsum 求和约定, 故省略了吸收规则关于哑指标 j 的求和操作 $\sum_j$, 求和后的维度与左端匹配.

$$R(Y^{l-1}) = R(Y^l)\odot\frac{D^l\odot\Delta Y^{l-1}}{\Delta Y^l} \tag{14}$$

图 2 相应地给出了传播过程的直观展示. 第 l 层相关性得分 $R(Y_j^l)$ 位于左列, 第 $l-1$ 层 $R(Y_i^{l-1})$ 位于上行. 相关性信息共有 $M\times N$ 个, 第 l 层相关性得分按行分解分配, 如黄色所示; 第 $l-1$ 层相关性得分由按列求和吸收而来, 如蓝色所示.

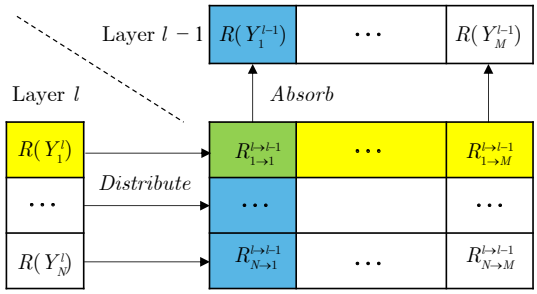


图 2　增量相关性的逐层分配与吸收

Fig. 2　Layer-wise distribution and absorption of the increment relevance

2.2　LID-Taylor 的本质

LID-Taylor 允许计算相关性的解析解, 因为各层参考输出 $\hat{Y}^l$ 是由参考点 $\hat{X}$ 唯一计算而来, 而不是任意选取. 从而由式 (14) 容易看出, 增量将在相关性传播过程中相互抵消. 与 DeepLIFT 类似, 对于类别 c, 将 LID-Taylor 方法的顶层相关性得分设为 $R(Y^L) = e_c\odot\Delta Y^L$; 根据逐层计算式 (14) 从后向前递推, 下一层相关性得分 $R(Y^{L-1}) = e_c D^L\Delta Y^{L-1}$, 下下一层的相关性得分 $R(Y^{L-2}) = e_c D^L D^{L-1}\Delta Y^{L-2}$. 由此容易归纳到式 (15), 可由顶层相关性 $R(Y^L)$ 直接计算出第 l 层的增量相关性得分, 其中 $G^{l+1} = D^L D^{L-1}\cdots D^{l+1}$ 满足链式法则.

$$R(Y^l) = e_c\odot G^{l+1}\odot\Delta Y^l \tag{15}$$

而由式 (12), 相应类别 c 的决策增量 Δf 关于 Y^l 的泰勒展开可如式 (16) 所示. 显然, 梯度 $\frac{\partial Y_c^L}{\partial Y^l} = e_c G^{l+1}$, 因此 LID-Taylor 的增量相关性 $R(Y^l)$ 与决策增量 Δf 的一阶泰勒展开 $\frac{\partial Y_c^L}{\partial Y^l}\odot\Delta Y^l$ 等价.

$$\Delta f = \frac{\partial Y_c^L}{\partial Y^l}\odot\Delta Y^l \tag{16}$$

上述等价结果揭示, LID-Taylor 的增量相关性在本质上正是梯度 $\frac{\partial Y_c^L}{\partial Y^l}$ 与增量 ΔY^l 的乘积.

2.3　逐层梯度传播

根据上述讨论, LID-Taylor 的增量相关性计算将由式 (14) 转化为式 (15). 其中, 式 (15) 的增量 ΔY^l 可由如式 (12) 的前向逐层传播计算得来, 而此处的梯度 G^{l+1} 或 $\frac{\partial Y_c^L}{\partial Y^l}$ 同样可由式 (17) 反向逐层传播计算得来, 其中, $G^L = D^L$, $l = L-1$, $L-2$, $\cdots$, 1.

$$G^l = G^{l+1}\odot D^l \tag{17}$$

综上所述, LID-Taylor 的增量相关性计算实际可转换为如下两步法: 1) 式 (17) 的梯度逐层传播. 2) 式 (15) 的相关性逐层传播. 因此, 与已有神经元相关性解释方法不同, LID-Taylor 的相关性计算由式 (14) 转化为式 (15), 可有效避免 DeepLIFT、LRP-0 等所涉及的除法运算, 不仅能够提升相关性计算的数值稳定性, 而且也为神经元的相关性解释提供了一种新视角.

因此 LID-Taylor 利用逐层统一的泰勒分解, 探究了相关性分配与后向传播的深刻联系.

2.4　与 LRP-0 的关系

LID-Taylor 对所有层采用统一的泰勒规则计算神经元相关性. 而 LRP-0、DeepLIFT 等方法对最大池化、激活等非线性层设计了其他计算规则. 例如对于 ReLU 层, LRP-0 直接跳过 (Pass) 神经元的相关性计算, 逐层相关性传播复用的是前层信息.

为了单独分析增量的作用, 将泰勒展开的思想也引入 LRP-0, 仿照 LID-Taylor, 将式 (6) 的 W^l 替换为 D^l, 得到类比推广方法 LRP-Taylor. 类似于式 (15), 当 LRP-Taylor 统一应用于所有层时, 相关性计算可由式 (18) 完成.

$$R(Y^l) = e_c\odot G^{l+1}\odot Y^l \tag{18}$$

对比式 (15) 与式 (18), 可以看出二者的不同.

假设对相同的顶层相关性分别应用 LRP-Taylor 与 LID-Taylor 规则, 前者将相关性分配给净输出 Y^l, 而后者则是 ΔY^l. 显然, 当参考输出 $\hat{Y}^l=0$ 时, 两者的相关性解释等价. 但参考输出 $\hat{Y}^l$ 通常不能满足为零. 当分配出现差异时, 可认为 LRP-Taylor 上层的一部分相关性被 $\hat{Y}^l$ 吸收, 产生了一部分错误归因. 通过引入增量, LID-Taylor 可以实现更为准确的相关性计算. 因此, LRP-Taylor 是 LID-Taylor 的特殊情况. 更宽泛地说, 增量相关性是相关性的推广.

3 逐层增量分解的提升方法

3.1 对比提升

由第 1.5 节可知, 以 $f(X)=Y_c^L=Z_c$ 作为决策函数的 LID-Taylor 还不具有类区分性. SG-LRP 作为 LRP 的对比提升, 虽然通过修改 LRP 的顶层相关性设置, 改善了类区分性, 但直接设置顶层相关性缺乏解释机理的严谨性, 并未明确给出决策函数. 同时, 如第 1.5 节所述的, SG-LRP 将非目标类全部设置为负相关性同样存在缺陷.

基于上述考虑, 本节进一步探讨 LID-Taylor 的对比改进. 根据第 2.2 节, 增量相关性是梯度与增量的乘积, 顶层相关性也应遵循该规律. LID-Taylor 满足上述规律, 因为 $f(X)=Z_c$, 且 $R(Y^L)=e_c\Delta Z$, 其中 $\frac{\partial f}{\partial Z}=e_c$. 而 SG-LRP 使用式 (10) 所述的梯度作为顶层相关性显然不符合相关性的本质属性.

因此, 本节给出一种具有类区分性的对比方法 ST-LID-Taylor (Softmax Taylor, ST). 具体而言, 以 Softmax 的第 c 类输出 P_c 作为决策函数 $f(X)=P_c$, 继而将决策增量 ΔP_c 分解为顶层相关性之和. 类比式 (15), (16), 顶层相关性设置为 ST (Softmax 的一阶泰勒展开), 即 $R(Y^L)=P_c'\Delta Z$, 其中 ΔZ 是 logits 层输出增量, $P_c'=\frac{\partial P_c}{\partial Z}$ 与式 (10) 相同, 而 ST-LID-Taylor 的中间层相关性计算与 LID-Taylor 保持相同.

ST-LID-Taylor 也具有类区分性, 通过顶层后的非线性激活函数 Softmax, 考虑了不同类别的互斥关系. 另外, 还利用增量 ΔZ 的正负性修正了相关性的准确性. 对于非目标类 $i\neq c$ 而言, 当得分 Z_i 上升时, $\Delta Z_i>0$, $\frac{\partial P_c}{\partial Z_i}<0$, 故类别 i 应设置为负贡献; 反之, 当得分下降时, 类别 i 应设置为正贡献. 特别地, 当令 $\Delta Z=Z$, 可自然引入 ST-LRP-0 方法作为 ST-LID-Taylor 方法的特例, 以对照分析相对于 SG-LRP-0 (中间层利用 LRP-0 规则) 的性能提升.

要指出的是, ST-LID-Taylor 的顶层相关性依然涉及到梯度项 $\frac{\partial P_c}{\partial Z}$, 因而 ST-LID-Taylor 同样存在类似 SG-LRP 的梯度饱和问题.

3.2 非线性提升

从另一个角度看, LID-Taylor 的近似误差主要源于式 (12) 中的高阶项 ϵ, 而实际网络结构存在大量非线性运算层, 因而第 2.1 节的线性泰勒分解难免带来解释误差. 事实上, 不仅是 Softmax 层, LID-Taylor 在所有非线性层都可能存在着梯度饱和问题, 例如 ReLU 层在输入小于零的情况下产生梯度弥散.

根据梯度定理[24], 任意路径上函数梯度的向量积分总是等于函数增量, 则神经元 Y_j^l 关于神经元 Y_i^{l-1} 的增量 ΔY_j^l 可如式 (19) 所示, 其中, $\hat{Y}^{l-1}\rightarrow Y^{l-1}$ 代表任意积分路径, 通常选择为直线. 因此, 相较于式 (12), 使用积分计算神经元的增量将不受高阶项 ϵ 的影响.

$$\Delta Y_j^l=\sum_i\int_{\hat{Y}^{l-1}\rightarrow Y^{l-1}}D_{ji}^l\mathrm{d}Y_i^{l-1} \tag{19}$$

基于上述考虑, 本节引入一种基于梯度积分的 LID-Taylor 非线性提升方法 LID-IG. 具体而言, LID-IG 使用梯度积分计算相关性信息, 以 $\int D_{ji}^l\mathrm{d}Y_i^{l-1}$ 替换式 (12) 中的 $D_{ji}^l\Delta Y_i^{l-1}$, 其他计算过程保持不变. 从而, 类似式 (14), LID-IG 的矩阵−向量化表达如式 (20) 所示, 此处的积分结果为矩阵.

$$R(Y^{l-1})=\frac{R(Y^l)}{\Delta Y^l}\odot\int_{\hat{Y}^{l-1}\rightarrow Y^{l-1}}D^l\odot\mathrm{d}Y^{l-1} \tag{20}$$

值得注意的是, 虽然式 (19) 中的梯度积分之和与路径的选择无关, 但梯度积分各个分量随着路径的选择会发生变化, 而产生不同的解释. LID-IG 方法通过逐层设计路径区别于 IG. 具体而言, IG 将参考点到输入确定为积分路径后, 中间层梯度计算涉及的路径随之完全确定; 而根据每层的参考输出 $\hat{Y}^{l-1}$ 与输出 Y^{l-1}, LID-IG 方法原则上可逐层规划任意积分路径, 还可以通过平滑梯度减轻噪声的影响.

式 (19) 的梯度积分可通过离散采样具体计算. 如, 在积分路径上选择等间隔的 n 个采样点 $Y_i^{l-1}(k)=\hat{Y}_i^{l-1}+(k/n)\Delta Y_i^{l-1}$, $k=1,\cdots,n$, 可将积分 $\int D_{ji}^l\times\mathrm{d}Y_i^{l-1}$ 简化为乘积 $\bar{D}_{ji}^l\Delta Y_i^{l-1}$, 其中 $\bar{D}_{ji}^l=(1/n)\sum_k D_{ji}^l(k)$ 表示平均局部偏导数, $D_{ji}^l(k)$ 表示在 $Y_i^{l-1}(k)$ 处的采样, 进而可引入平均雅可比矩阵 $\bar{D}^l$ 实现增量 ΔY^l 的逐层计算, 式 (21) 给出了计算过程的矩阵−向量化表达

$$\Delta Y^l=\bar{D}^l\odot\Delta Y^{l-1} \tag{21}$$

从而, 类似式 (17) 和式 (15), LID-IG 的增量相关性计算可利用如下两步法完成: 1) 式 (22) 的平均梯度逐层传播. 2) 式 (23) 的逐层相关性传播.

$$\bar{G}^l = \bar{G}^{l+1} \odot \bar{D}^l \tag{22}$$

$$R(Y^l) = e_c \odot \bar{G}^{l+1} \odot \Delta Y^l \tag{23}$$

类似地, 平均梯度 $\bar{G}^{l+1} = \bar{D}^L \bar{D}^{L-1} \cdots \bar{D}^{l+1}$ 满足链式法则, 其中 $l = L-1,\ L-2,\ \cdots,\ 1$, 而 $\bar{G}^L = \bar{D}^L$. 从平均梯度的计算方式而言, LID-IG 可在一定程度上缓解 ST-LID-Taylor 的梯度饱和问题.

3.3　交叉组合提升

本节进一步探讨非线性与对比提升的交叉组合策略. 首先对前文涉及的方法作如下总结.

1) 第 2.1 节的 LID-Taylor 作为 LRP-0 的增量与泰勒推广, 依然是面向中间层的相关性计算规则, 提出了线性层 (全连接层、卷积层) 与非线性层 (ReLU、MaxPool 层) 的统一计算方法.

2) 第 3.1 节的对比提升策略为顶层带来了新变化. 通过利用非线性激活函数 Softmax, 构造类别互斥的顶层相关性, 得到 ST-LID-Taylor 方法.

3) 第 3.2 节的非线性提升策略替换了一阶泰勒展开的思想, 为逐层相关性计算引入积分梯度运算, 并率先应用于中间层, 得到 LID-IG 方法.

上述方法的设计思想是相互独立的, 并且允许进一步组合. 例如, 当对 ST-LID-Taylor 在顶层引入非线性提升, 即得到交叉组合方法 SIG-LID-Taylor (Softmax integrated gradients, SIG). 通过将顶层相关性设置为 SIG, 即 $R(Y^L) = \bar{P}'_c \Delta Z$, 其中 $\bar{P}'_c$ 是 c 类 Softmax 概率的平均梯度 (参考 $\bar{D}^l_{ji}$), 从而可缓解 Softmax 梯度饱和问题. 在 SIG-LID-Taylor 基础上, 通过进一步引入中间层的非线性提升, 可最终得到本文的交叉组合方法 SIG-LID-IG. 根据 SIG 与式 (23), 容易得出 SIG-LID-IG 的任意层相关性计算如式 (24) 所示. 算法 1 给出了 SIG-LID-IG 的伪代码.

$$R(Y^l) = \bar{P}'_c \odot \bar{G}^{l+1} \odot \Delta Y^l \tag{24}$$

算法 1. SIG-LID-IG

1) 输入图像 X, 指定类别 c, 网络模型 f, 参考输入 $\hat{X}$, 积分步数 n;
2) For $l = 1$ to L;
3) 计算各层输出 Y^l, 参考输出 $\hat{Y}^l$ 与增量 ΔY^l;
4) 计算 Softmax 的平均梯度 $\bar{P}'_c$;
5) 根据式 (24) 计算顶层相关性 $R(Y^L) = \bar{P}'_c \odot \Delta Y^L$;
6) End for;
7) For $l = L-1$ to 0;
8) 根据式 (22) 计算平均梯度 $\bar{G}^{l+1}$;
9) 根据式 (24) 计算相关性 $R(Y^l) = \bar{P}'_c \odot \bar{G}^{l+1} \odot \Delta Y^l$;
10) End for;
11) 输出所有层的相关性 $R(Y^l)$, $l \in \{0,\ 1,\ \cdots,\ L\}$.

与 LID-Taylor 相比, SIG-LID-IG 的算法复杂度主要增加了两部分. LID-Taylor 的算法复杂度等同于一次网络运行. 积分梯度的引入使得计算量变为原来的 n 倍, 在实验中 $n = 10$. 而顶层相关性带来的变化仅为网络添加了额外一层的计算量. LID-Taylor 在 VGG-16 模型上生成热力图的运行时间大约是 23 ms, 而 SIG-LID-IG 则是 64 ms, 仅约为前者的 3 倍.

最后, 对本文涉及到的所有方法进行综合比较. 表 1 展示了在实际应用中部分解释方法的逐层规则细节, 例如 LRP-0 解释方法指的是在线性层上使用该规则, 而非线性层则不同, 对于 ReLU 层使用跳过计算 (Pass), 对于 Maxpool 层使用 WTA 规则; LID-Taylor 采用了统一的计算方式. 需要注意的是, SIG-LID-IG 在线性层的规则表示为 LID-Taylor*, 这是因为当遇到线性层, LID-IG 将退化为 LID-Taylor, 二者相互等价.

表 1　不同方法的逐层规则对比
Table 1　Layer-wise rule comparison of different methods

方法名	LRP-0	LID-Taylor
顶层	$e_c \odot Z$	$e_c \odot \Delta Z$
线性层	LRP-0	LID-Taylor
非线性层	Pass, WTA	LID-Taylor
方法名	ST-LID-Taylor	SIG-LID-IG
顶层	ST	SIG
线性层	LID-Taylor	LID-Taylor*
非线性层	LID-Taylor	LID-IG

为进一步详细比较, 表 2 与表 3 展示了部分方法的顶层相关性与中间层计算公式对比. 除了文中涉及的组合方法, 对其他可能的组合进行以下补充说明: 1) LID-Taylor 对顶层仅进行非线性提升而无对比提升是无效的, 这是因为非线性提升针对的是非线性函数, 而 logits 层是线性的. 因此仅在对比提升引入了非线性函数 Softmax 之后的非线性提升才有效. 2) 属于中间层的线性层无需非线性提升, 在表 1 的 SIG-LID-IG 一栏中以 LID-Taylor* 明确指出, 这可以减少约一半的额外计算量. 3) 不利用增量的非线性提升存在一部分无效情况, 例如

表 2 顶层相关性对比
Table 2 Comparison of top layer relevance

方法名	顶层相关性
LRP-0	$e_c \odot Z$
LID-Taylor	$e_c \odot \Delta Z$
SG-LRP	P'_c
ST-LID-Taylor	$P'_c \odot \Delta Z$
SIG-LID-IG	$\bar{P}'_c \odot \Delta Z$

表 3 中间层规则对比
Table 3 Comparison of middle layer rule

方法名	相关性计算规则
LRP-0	$R(Y^{l-1}) = \frac{R(Y^l)}{Y^l} \odot W^l \odot Y^{l-1}$
DeepLIFT	$R(Y^{l-1}) = \frac{R(Y^l)}{\Delta Y^l} \odot W^l \odot \Delta Y^{l-1}$
LID-Taylor	$R(Y^{l-1}) = \frac{R(Y^l)}{\Delta Y^l} \odot D^l \odot \Delta Y^{l-1}$
LID-IG	$R(Y^{l-1}) = \frac{R(Y^l)}{\Delta Y^l} \odot \bar{D}^l \odot \Delta Y^{l-1}$

ReLU 函数属于分段线性函数, 这时按照第 2.4 节的描述, 中间层的参考等价为零 $\hat{Y}^l = 0$, 而函数的正、负半平面都是线性的, 因此提升无效.

利用多层相关性可以进一步丰富热力图的语义信息. 受到 LayerCAM[12] 多尺度热力图方法的启发, 本文还将在 SIG-LID-IG 方法的基础上进一步引入多尺度的热力图, 对不同层的热力图拉伸并叠加, 融合纹理、物体等多尺度特征, 提高热力图的语义信息, 以更合理地评估神经元相关性解释方法的准确性.

综上所述, 图 3 在增量 (Increment)、对比 (Contrast)、非线性 (Nonlinearity) 三个维度下, 给出本文的神经元相关性解释方法的全景图. 其中, SIG-LID-IG 是本文方法在三个维度下的集大成者. 此外, 图 3 还给出了 LRP 系列下的 LRP-0、SG-LRP, 作为本文提出方法的对照. 值得一提的是, 上述交叉组合方法均针对的是 LID-Taylor. 由第 3.1 节知, 无参考点的 ST-LRP 可作为 ST-LID-Taylor 的特例.

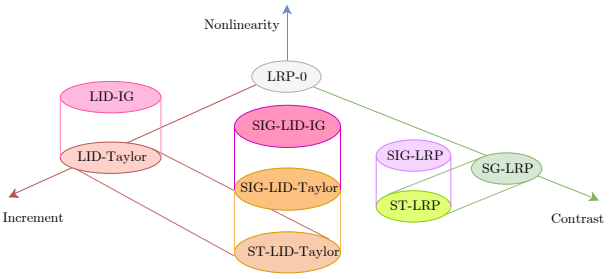


图 3 本文神经元相关性解释方法的全景图
Fig. 3 Panorama of the neuron relevance explanation methods in this article

因此, 如果对 ST-LRP 顶层神经元相关性引入第 3.2 节的非线性提升, 可得到 SIG-LRP 作为对照.

4 实验结果与分析

本文所有实验均在 PyTorch 框架下完成. 使用的数据集为样本数量为 5 000 的 ImageNet 验证集 Val 的子集. 神经网络模型选择为在 ImageNet 上预训练的 VGG-16[2].

本文主要以热力图定性、定量评估不同的解释方法. 为了指示热力图所对应的网络层, 此处引入一种所谓的阶段简记法. 为此, 将 VGG-16 网络结构划分为如下 5 个阶段: s1、s2、s3、s4、s5 (分别对应卷积层 Conv1_2、Conv2_2、Conv3_3、Conv4_3、Conv5_3 之后的最大池化层), 从而可在不同解释方法的名称之后添加阶段标识作为后缀, 以具体指示不同阶段下的解释方法及对应的热力图结果. 如, SIG-LID-IG-s5 指代第 5 阶段下的 SIG-LID-IG 方法及对应单尺度热力图结果; 而 SIG-LID-IG-s54321 指代全部 5 个阶段下的 SIG-LID-IG 方法对应的多尺度热力图.

除了图 3 给出的本文系列方法外, 本节实验的比较方法还包括: GradCAM[11]、LayerCAM[12]、ScoreCAM[13]、IG[21]、LRP-0[14]、SG-LRP-ZP[20] 等基准解释方法. 其中, 对于本文直接涉及到 LRP 的方法, 在实现时均融合了 LRP-ϵ 修正与 LRP-0 的相关性计算规则, 包括 SG-LRP-0、ST-LRP-0、SIG-LRP-0, 与 SG-LRP-ZP 一同在实验中进行对比分析.

4.1 单尺度与多尺度热力图

图 4 具体给出了四组样本的 SIG-LID-IG-s54321 热力图示例. 热力图是一种常见的卷积层贡献度可视化方法[10], 通过将贡献度按通道叠加, 拉伸至原图像尺寸, 观察不同空间区域的贡献度大小. 其中, 每组样本的前景均含有 A、B 两个类别. 第一组的是牛奏与虎猫, 第二组的是斑马与大象, 第三组的是孔雀与公鸡, 第四组的是蓝鹀与金翅雀. 对于每组样本实验, 第一行给出原始图片, 第二、三行分别是对应类别 A、B 的决策函数 $f(X)$ 的热力图. 其中, 红色代表热力图的空间语义与决策函数 $f(X)$ 正相关, 而蓝色表示负相关, 白色表示不相关. 结果表明, 无论决策函数面向 A 或 B, SIG-LID-IG-s54321 热力图的正相关性区域均可精准定位目标, 还具有一定的局部聚焦能力与细节刻画能力. 同时, 对于 VGG-16 网络, 类别 A 的预测概率均高于类别 B, 即 VGG-16 对四组样本的预测类别均为 A. 当类别 A 在粗粒度上显著区别于类别 B 时, 如前三组样本, 热力图将以归因类别 A 的正相关性为主; 当类别

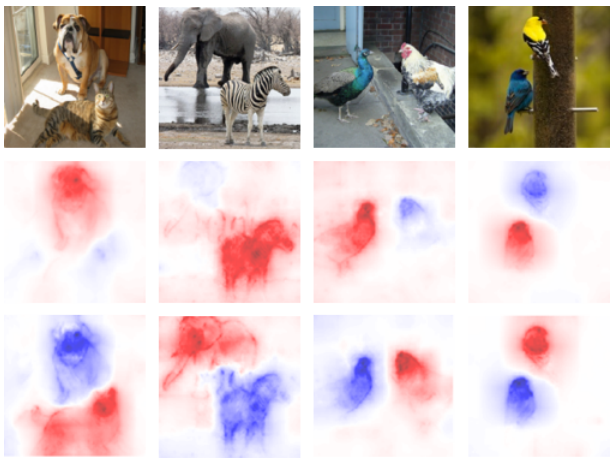


图 4　SIG-LID-IG-s54321 多尺度热力图

Fig. 4　Multi-scale heatmaps of SIG-LID-IG-s54321

A 在细粒度上显著区别于类别 B 时, 如第四组的蓝鸦与金翅雀, 热力图在归因类别 A 的正相关性的同时, 还体现了类别 B 的负相关性. 而对于第三行, 决策函数 $f(X)$ 面向的是类别 B, 热力图在归因类别 B 的正相关性的同时, 还体现了类别 A 的负相关性, 此时均衡的正、负相关性表现符合预期. 实验结果支持了第 3.1 节的观点, 即引入顶层相关性可以显著提高热力图的类区分度, 提升决策归因的精确性.

为了进一步直观理解本文方法, 针对图 4 第一组示例, 图 5 和图 6 还展示了 SIG-LID-Taylor、LID-IG 和 SIG-LID-IG 在 s1、s2、s3、s4、s5 不同阶段下的单尺度热力图结果, 分别对应类别 A、B 的决策函数. 根据各阶段的热力图可知, SIG-LID-IG 均取得明显优于 SIG-LID-Taylor 的相关性决策归因. 同时注意到, SIG-LID-IG 的低阶段热力图具有精准的细节, 而 SIG-LID-Taylor 的热力图充满杂乱无章的噪声, 直观验证了非线性增量相关性计算对于浅层特征归因的精确性与鲁棒性. 此外, 在类区分性表现上, SIG-LID-IG 相较于 LID-IG, SIG-LID-Taylor 的优势同样明显. 实验证实了第 3.2 节的观点, 即采用积分梯度可以缓解中间层的梯度消失现象, 提高决策归因在面向中间层特征的鲁棒性.

综上可知, 本文 SIG-LID-IG 方法在不同尺度下的空间语义与决策函数均可有效适配, 实现了更加精确、鲁棒的决策归因.

4.2　Probability Change

受 Average drop[13] 的启发, 本节通过类别 A 的预测概率变化准则 (Probability change, PC) 来具体评价热力图的精确性. 具体地, 先将热力图转化为二值化掩膜, 然后分析原始图像叠加掩膜前后

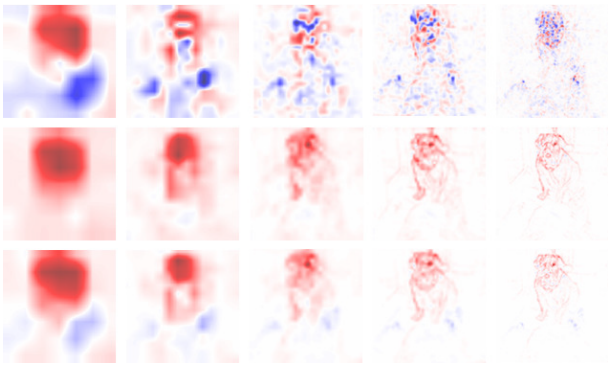


图 5　决策类别为牛獒时不同解释方法单尺度热力图展示 (从左至右: s5、s4、s3、s2、s1; 从上到下: SIG-LID-Taylor、LID-IG、SIG-LID-IG)

Fig. 5　Single-scale heatmaps of different explanation methods when the decision category is bull mastiff (From left to right: s5, s4, s3, s2, s1; From top to bottom: SIG-LID-Taylor, LID-IG, SIG-LID-IG)

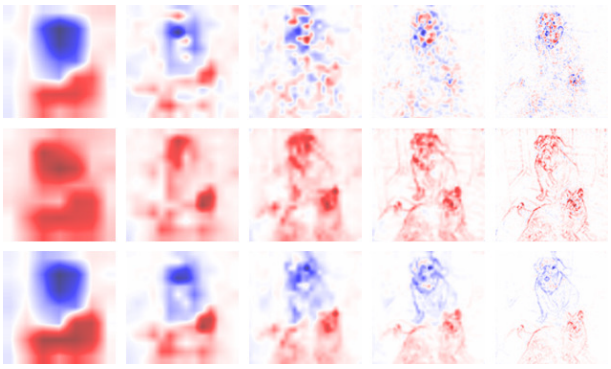


图 6　决策类别为虎猫时不同解释方法单尺度热力图展示 (从左至右: s5、s4、s3、s2、s1; 从上到下: SIG-LID-Taylor、LID-IG、SIG-LID-IG)

Fig. 6　Single-scale heatmaps of different explanation methods when the decision category is tiger cat (From left to right: s5, s4, s3, s2, s1; From top to bottom: SIG-LID-Taylor, LID-IG, SIG-LID-IG)

模型预测的概率变化, 以此量化热力图空间语义相关性的准确性. 实验中, 所有方法的热力图掩膜保留相同比例的像素, 且全部对应第 5 阶段的单尺度热力图. 图 7 给出了 GradCAM、LayerCAM、ScoreCAM、IG、LRP-0、SG-LRP-0、SG-LRP-ZP 以及本文方法 ST-LRP-0、SIG-LRP-0、SIG-LID-Taylor、LID-IG、SIG-LID-IG 的 s5 阶段热力图结果. 图 8 给出了不同方法的预测概率变化曲线图, 横轴表示热力图掩膜保留像素的比例, 而纵轴表示模型叠加掩膜后产生的预测概率变化.

由图 7 可知, 现有工作中的三种方法 GradCAM、LayerCAM、ScoreCAM 明确不具有类区分性的解释能力. 由图 8 可知, ScoreCAM 与 IG 取得

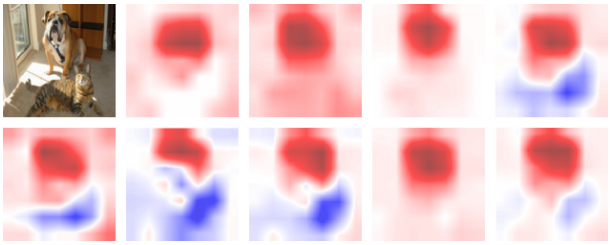


图 7 不同方法的第 5 阶段单尺度热力图对比 (从左至右: GradCAM、LayerCAM、ScoreCAM、IG、LRP-0、SG-LRP-ZP、SIG-LID-Taylor、LID-IG、SIG-LID-IG)

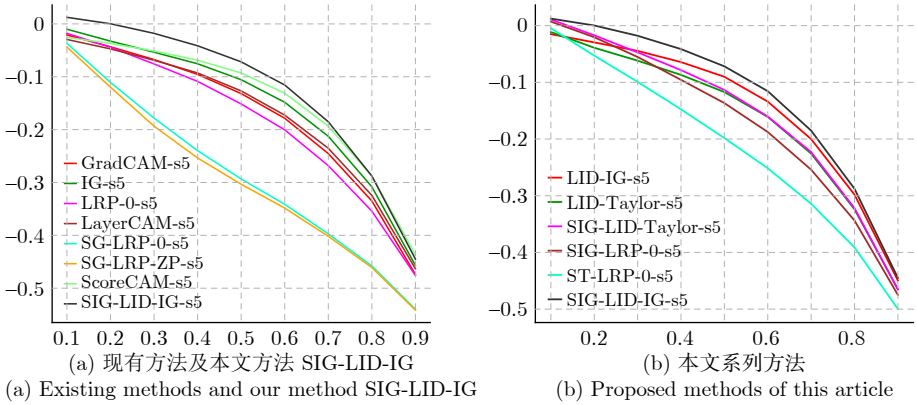
Fig. 7 Comparison of stage 5 single-scale heatmaps for different methods (From left to right: GradCAM, LayerCAM, ScoreCAM, IG, LRP-0, SG-LRP-ZP, SIG-LID-Taylor, LID-IG, SIG-LID-IG)

了相对较好的综合性能, LayerCAM、GradCAM、LRP-0 次之. 而令人意外的是, 相较于 LRP-0, 相应的对比提升方法 SG-LRP-0 与 SG-LRP-ZP 在 PC 准则下性能下降较为明显, 表明该方法在实现类区分性解释上具有潜在的缺陷, 由图 7 所示的 SG-LRP-ZP 热力图结果可见一斑. 同时, 图 8 左图清楚展示了现有工作相较于本文组合方法 SIG-LID-IG 的劣势, 而图 8 右图则展示了本文系列方法的性能表现. 一方面, ST-LRP-0 方法明显优于 SG-LRP-0, 且进一步引入了非线性提升的 SIG-LRP-0 方法优于 LRP-0. 而另一方面, 虽然基于增量的 LID-Taylor 方法同样优于 LRP-0, 但还不足以媲美 ScoreCAM 和 IG. 不过, 对 LID-Taylor 引入中间层非线性提升后, 所得的 LID-IG 方法已可媲美 ScoreCAM 和 IG, 验证了中间层非线性提升的有效性. 进一步地, 当对 LID-IG 的顶层引入对比与非线性提升后, 本文的最终方法 SIG-LID-IG 实现了性能上的跃升, 在所有比较方法中取得了最佳的性能指标. 特别地, 由图 8 右图可知, SIG-LID-IG 允许原始样本移除高达 20% 的像素而不损害 VGG-16 的预测精度; 而移除不多于 20% 的像素后, VGG-16 的预测概率反而增加, 恰恰体现了负相关性的作用, 也即移除负相关性像素能够改善网络决策, 从而合理验证了引入顶层对比提升策略对于精确实现类区分性的准确性. 要指出的是, 相对于 LID-Taylor, 虽然 SIG-LID-Taylor 同样受益于该提升策略, 但是随着像素移除量的增加, SIG-LID-Taylor 逐渐趋于 LID-Taylor, 再次彰显了 SIG-LID-IG 引入中间层非线性提升改善热力图相关性决策归因的必要性和准确性. 为清晰起见, 表 4 给出了 ScoreCAM、IG 以及本文 SIG-LID-Taylor、LID-IG、SIG-LID-IG 的具体 PC 数值.

此外, 图 9 给出多尺度解释方法 LayerCAM 与本文组合方法 SIG-LID-IG 在不同多尺度热力图下的 PC 性能分析. 对应类别 A 的决策函数 $f(X)$, 图 10 给出图 4 的相同四组样本下的 LayerCAM-s54321、SIG-LID-IG-s54321 多尺度热力图结果. 由图 9 可知, 一方面, 两方法在多尺度热力图下的 PC 性能均优于单尺度的情形, 清楚地展示了二者对于多尺度决策归因的有效性; 另一方面, 本文方法 SIG-LID-IG 在不同尺度组合的敏感性上要明显弱于 LayerCAM, 也进一步验证了 SIG-LID-IG 在单尺度决策归因上相较于 LayerCAM 的精确性. 事实上, 由图 10 的多尺度热力图结果明显可知, 对于类别 A 的决策归因, LayerCAM 在精确性与合理性上均不及本文方法 SIG-LID-IG.

4.3 Minimal Patch

神经网络作为黑箱, 内部神经元的复杂组合将严重阻碍人们关于网络决策的可信分析与理解. 自然地, 内部不同神经元对最终决策可能产生不同的正、负相关性归因. 第 4.2 节主要通过类别 A 的预测概率下降评价热力图的合理性, 主要源于热力图



(a) 现有方法及本文方法 SIG-LID-IG

(a) Existing methods and our method SIG-LID-IG

(b) 本文系列方法

(b) Proposed methods of this article

图 8 不同方法的 PC 评估折线图

Fig. 8 Line chart of PC evaluation of different methods

表 4　本文方法 SIG-LID-Taylor、LID-IG、SIG-LID-IG 与 ScoreCAM、IG 的 PC 实验数值比较

Table 4　Comparison of PC experimental values between the proposed methods SIG-LID-Taylor, LID-IG, SIG-LID-IG and ScoreCAM, IG

比例	IG	ScoreCAM	SIG-LID-Taylor	LID-IG	SIG-LID-IG
0.1	−0.010 06	−0.025 44	0.010 88	−0.015 03	0.012 43
0.2	−0.032 59	−0.037 00	−0.017 05	−0.029 54	0.000 25
0.3	−0.053 58	−0.051 20	−0.047 14	−0.044 54	−0.017 88
0.4	−0.075 42	−0.068 61	−0.078 46	−0.064 28	−0.041 39
0.5	−0.105 52	−0.092 72	−0.113 35	−0.090 19	−0.071 76
0.6	−0.148 30	−0.131 05	−0.160 32	−0.133 97	−0.115 59
0.7	−0.212 55	−0.193 31	−0.222 92	−0.200 09	−0.185 15
0.8	−0.308 24	−0.287 81	−0.321 89	−0.298 01	−0.287 73
0.9	−0.456 71	−0.436 46	−0.465 28	−0.449 57	−0.444 88

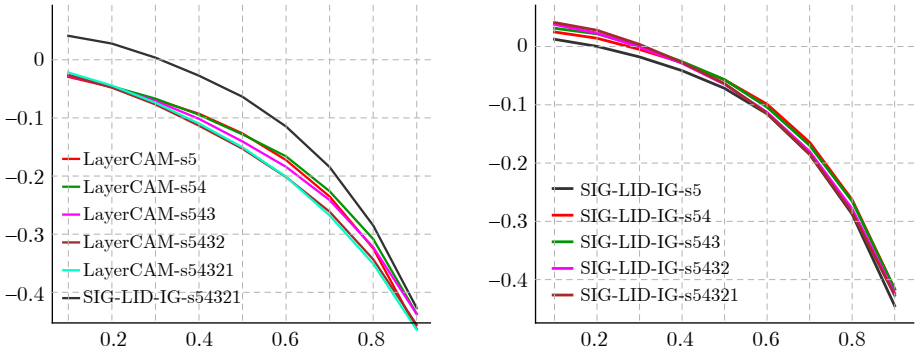
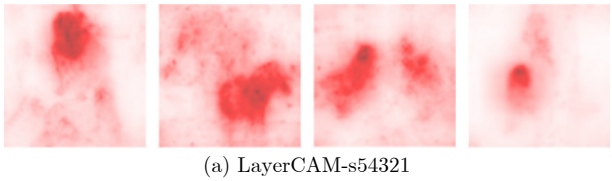
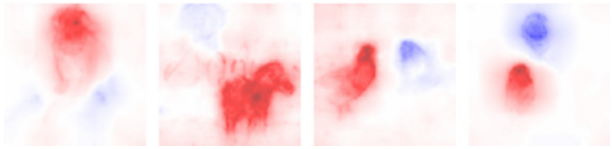


图 9　LayerCAM 与 SIG-LID-IG 在不同多尺度热力图下的 PC 评估折线图

Fig. 9　Line chart of PC evaluation between LayerCAM and SIG-LID-IG with different multi-scale heatmaps



(a) LayerCAM-s54321



(b) SIG-LID-IG-s54321

图 10　多尺度热力图

Fig. 10　Multi-scale heatmaps

正相关性区域的掩膜丢弃. 进一步地, 本节将特别从神经元的负相关性作用去评估不同解释方法的有效性与合理性.

作为负相关性解释的代表性方法, SG-LRP-ZP 通过最大补丁 (Maximal patch) 策略评估热力图显著性区域的定位精度. 具体而言, 该策略以方形掩膜移除热力图最大正相关性区域, 遮挡正相关性区域自然导致预测概率的下降偏好, 因而预测概率下降大的方法更好. 作为第 4.2 节的补充, 本节对于类别 A 特别设计基于最小补丁 (Minimal patch) 的预测概率变化评估策略, 通过移除热力图最小点区域来具体分析各种解释方法的不同表现. 对于具有一定负相关性解释能力的方法, 最小点区域即为最大负相关性区域. 本节分别选择了对应不同半径 $r \in \{1, 2, 3, 5, 10, 20\}$ 的方形掩膜进行实验, 掩膜大小实际为 $2r+1$.

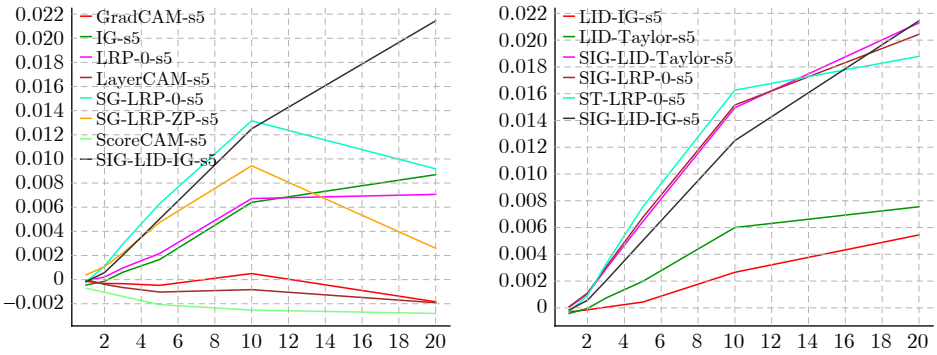
图 11 给出不同方法的预测概率变化情况, 横轴表示最小补丁的半径, 而纵轴表示叠加最小补丁后产生的预测概率变化. 类似地, 所有方法的热力图均对应 s5 阶段的单尺度热力图. 结果表明, 对于不具有负相关性解释能力的方法, 例如 ScoreCAM、LayerCAM、GradCAM, 掩膜最小补丁将自然导致预测得分下降. 注意到, 虽然 LRP-0、IG 以及本文方法 LID-Taylor、LID-IG 没有相应设计负相关性的解释机制, 掩膜最小补丁后预测得分却有所上升, 显示了上述方法的灵活性. 与上述几个方法相比, SG-LRP-0 与 SG-LRP-ZP 的负相关性解释能力相对更好, 但是当掩膜半径 r 大于 10 时, 预测得分即呈现下降趋势, 再次表明 SG-LRP 在实现类区分性

解释上具有潜在的缺陷. 而图 11 的左图展示了组合方法 SIG-LID-IG 相对于掩膜半径的准确性.

图 11 的右图展示了本文系列方法的性能表现. 在第 5 阶段下, SIG-LID-IG 明显优于 LID-Taylor、LID-IG, 但稍逊于 SIG-LID-Taylor 以及 ST-LRP-0、SIG-LRP-0. 尽管如此, 前文图 5 显示, SIG-LID-Taylor 的低阶段热力图存在杂乱无章的噪声, 不同尺度下决策归因的准确性明显不足, 而 ST-LRP-0、SIG-LRP-0 自然存在类似问题. 为更清楚地展示, 图 12 给出 ST-LRP-0、SIG-LRP-0、SIG-LID-Taylor、SIG-LID-IG 四种负相关性解释方法在不同多尺度热力图下的性能分析. 由结果可知, 虽然 SIG-LID-Taylor 等在半径 r 较小时, 得分略高于 SIG-LID-IG, 但随着层数的降低与半径增大, 性能迅速下降, 而 SIG-LID-IG 具有很好的精确性.

4.4 归因鲁棒性

在第 4.1 节的可视化实验中验证了 SIG-LID-IG 方法在中间层归因的精确性和鲁棒性, 第 4.2 节的定量实验进一步验证了其归因的准确性, 本节进

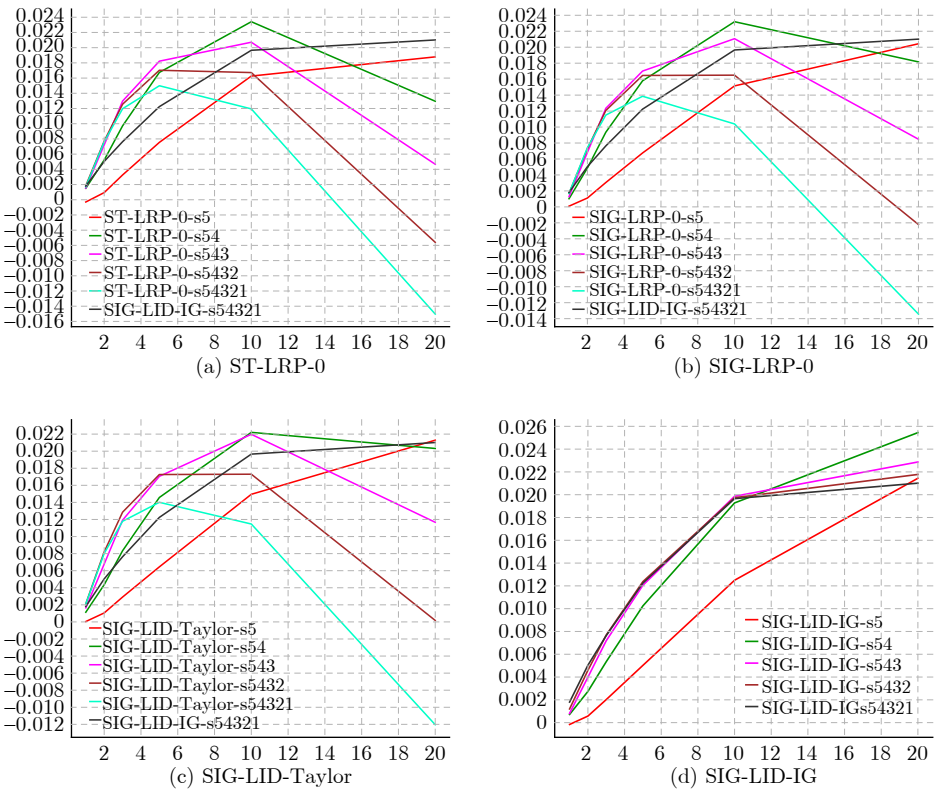


(a) 现有方法及本文方法 SIG-LID-IG (a) Existing methods and our method SIG-LID-IG

(b) 本文系列方法 (b) Proposed methods of this article

图 11 最小补丁热力图评估

Fig. 11 Minimal patch evaluation of heatmaps



(a) ST-LRP-0 (b) SIG-LRP-0

(c) SIG-LID-Taylor (d) SIG-LID-IG

图 12 多尺度热力图负相关性评价

Fig. 12 Negative relevance evaluation of multi-scale heatmaps

一步探究以 SIG 设置为顶层相关性的鲁棒性. 在文献 [20] 中指出, SG-LRP-ZP 在部分样本下存在归因不准确的情况, 当图像中的物体占据整个空间时, SG-LRP 可能会删除过多的区域. 图 13 选择了一些特殊的图像样本, 对 SG-LRP-ZP-s54321 和 SIG-LID-IG-s54321 的多尺度热力图进行比较. 结果表明, SG-LRP-ZP 热力图存在一定的问题, 例如将完整的钟表分为贡献度截然不同的两部分. 而 SIG-LID-IG 则可以正确归因, 体现了其精确性和鲁棒性.

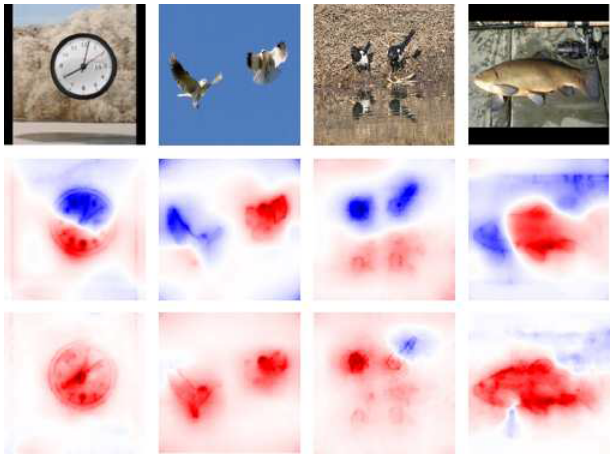


图 13　SG-LRP-ZP 和 SIG-LID-IG 的多尺度热力图对比

Fig. 13　Comparison of multi-scale heatmaps between SG-LRP-ZP and SIG-LID-IG

本文认为此现象产生的主要原因是 SG-LRP 顶层相关性的梯度饱和问题, 即当图像中的物体占据整个空间时, 目标类的特征明显, 而干扰特征较少, 因此模型的识别概率较高, 根据第 1.5 节描述, 这将产生 Softmax 函数的梯度饱和现象, 导致相关性数值非常小, 在逐层计算中产生不稳定性. 根据第 3.1 节理论所述, 顶层相关性 ST 依然存在这种现象, 而第 3.3 节的 SIG 通过引入积分梯度可以解决梯度饱和问题. 因此, 本节特别针对 ST-LID-Taylor 和 SIG-LID-Taylor 进行对比分析, 对图像添加平移缩放抖动, 观察热力图的稳定性, 以验证它们归因的鲁棒性.

图 14 给出了两组样本的热力图. 第一组位于左侧, 类别为大灰猫头鹰, 第二组位于右侧, 类别是墨西哥鲵. 每组图像都经过了随机平移缩放变换, 两次变换的结果展示为两列. 每张图像都与热力图进行了叠加, 以突出本实验的重点, 第一行热力图来自 ST-LID-Taylor 方法, 第二行来自 SIG-LID-Taylor 方法, 阶段为 s5. 实验结果符合预期, ST-LID-Taylor 在同组样本间的热力图出现了正、负相关性跳变的现象, 而其对应的非线性提升方法 SIG-LID-Taylor 实现了同组样本间热力图归因的精确性和鲁棒性. 实验证明了第 3.3 节的观点, 即在顶层上处理 Softmax 函数时, 引入非线性提升的顶层相关性 SIG 具有很好的鲁棒性.

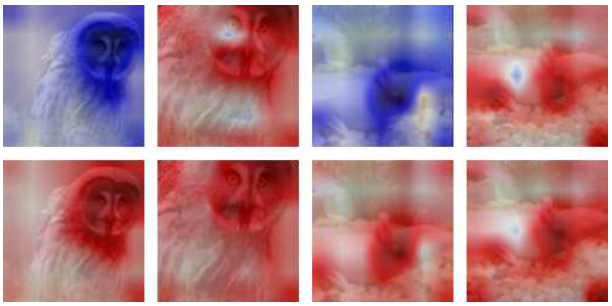


图 14　热力图的归因鲁棒性 (左, 右分别对应大灰猫头鹰和墨西哥鲵的两组平移缩放样本; 上、下分别为 ST-LID-Taylor 和 SIG-LID-Taylor 的热力图结果)

Fig. 14　Attribution robustness of heatmaps (Left and right correspond to the two groups of translation and scaling samples of great gray owl and Mexican salamander, respectively. The top and bottom show the heatmap results of ST-LID-Taylor and SIG-LID-Taylor, respectively)

5　总结

神经网络的黑箱特性严重阻碍了人们关于网络决策的直观分析与理解. 尽管文献报道了多种基于神经元贡献度分配的决策解释方法, 但是现有方法的解释一致性难以保证, 鲁棒性更是有待改进, 打开“黑箱”依然任重道远. 本文从神经元相关性概念入手, 通过对逐层神经元的贡献度得分的本原探索, 提出一种基于逐层增量分解的神经网络解释新方法 LID-Taylor, 且在此基础上先后引入针对顶层神经元相关性的对比提升策略, 以及针对所有层神经元相关性的非线性提升策略, 最后利用交叉组合策略得到本文最终方法 SIG-LID-IG, 实现了决策归因性能的鲁棒跃升. 在实验验证部分, 本文通过热力图对现有工作与本文方法的决策归因性能做了定性与定量评估. 结果显示, SIG-LID-IG 在神经元的正相关性、负相关性的决策归因合理性上均可媲美甚至优于现有工作. 相较于 LayerCAM, SIG-LID-IG 在多尺度热力图下同样取得了精确性更高、鲁棒性更强的决策归因. 本文后续工作将基于 SIG-LID-IG 这一解释方法探讨可迁移的神经网络黑箱攻击问题.

References

1　Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, **60**(6): 84–90

2　Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2014. 1–14

3 Szegedy C, Liu W, Jia Y Q, 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: IEEE, 2015. 1−9

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

5 Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(1): 142−158

6 Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1440−1448

7 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR, 2014.

8 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 818−833

9 Erhan D, Bengio Y, Courville A, Vincent P. Visualizing Higher-layer Features of a Deep Network, Technical Report 1341, University of Montreal, Canada, 2009.

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

11 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: IEEE, 2017. 618−626

12 Jiang P T, Zhang C B, Hou Q B, Cheng M M, Wei Y C. LayerCAM: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021, **30**: 5875−5888

13 Wang H F, Wang Z F, Du M N, Yang F, Zhang Z J, Ding S R, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE, 2020. 111−119

14 Bach S, Binder A, Montavon G, Klauschen F, Müller K R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 2015, **10**(7): Article No. e0130140

15 Montavon G, Binder A, Lapuschkin S, Samek W, Müller K R. Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, 2019. 193−209

16 Samek W, Montavon G, Lapuschkin S, AndersC J, Müller K R. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 2021, **109**(3): 247−278

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

18 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

19 Gu J D, Yang Y C, Tresp V. Understanding individual decisions of CNNs via contrastive backpropagation. In: Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer, 2018. 119−134

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

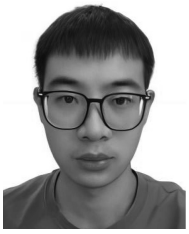
21 Sundararajan M, Taly A, Yan Q Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 3319−3328

22 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: Curran Associates Inc., 2017. 4768−4777

23 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 3145−3153

24 Colley S J. *Vector Calculus* (*Fourth edition*). Boston: Pearson, 2011.

陈艺元 南京邮电大学硕士研究生. 主要研究方向为深度学习模型的可解释性和迁移对抗攻击.
E-mail: cyy280113999@gmail.com
(**CHEN Yi-Yuan** Master student at Nanjing University of Posts and Telecommunications. His research interest covers interpretability of deep learning models and transferable adversarial attacks.)



李建威 南京邮电大学硕士研究生. 主要研究方向为深度学习模型的迁移对抗攻击.
E-mail: 1022010429@njupt.edu.cn
(**LI Jian-Wei** Master student at Nanjing University of Posts and Telecommunications. His research interest covers transferable adversarial attacks on deep learning models.)

邵文泽 南京邮电大学教授. 主要研究方向为计算成像, 视觉感知, 黑箱优化和可理解人工智能. 本文通信作者. E-mail: shaowenze@njupt.edu.cn
(**SHAO Wen-Ze** Professor at Nanjing University of Posts and Telecommunications. His research interest covers computational imaging, visual perception, black-box optimization, and understandable artificial intelligence. Corresponding author of this paper.)

孙玉宝 南京信息工程大学教授. 主要研究方向为计算机视觉, 快照压缩成像, 深度学习.
E-mail: sunyb@nuist.edu.cn
(**SUN Yu-Bao** Professor at Nanjing University of Information Science and Technology. His research interest covers computer vision, snapshot compressed imaging, and deep learning.)