计算机科学
COMPUTER SCIENCE

一种语义引导的神经网络关键数据路由路径算法

朱富坤, 滕臻, 邵文泽, 葛琦, 孙玉宝

引用本文

朱富坤, 滕臻, 邵文泽, 葛琦, 孙玉宝. 一种语义引导的神经网络关键数据路由路径算法[J]. 计算机科学, 2024, 51(9): 155-161.

ZHU Fukun, TENG Zhen, SHAO Wenze, GE Qi, SUN Yubao. Semantic-guided Neural Network Critical Data Routing Path [J]. Computer Science, 2024, 51(9): 155-161.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900109

一种语义引导的神经网络关键数据路由路径算法

朱富坤[1]　滕　臻[2]　邵文泽[1]　葛　琦[1]　孙玉宝[3]
1 南京邮电大学通信与信息工程学院　南京 210003
2 南京邮电大学贝尔英才学院　南京 210042
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
(zhufukun1999@163.com)

摘　要　近年来,由于人工智能在各领域的普及,研究神经网络的可解释方法及理解神经网络的运作机理已经成为一个愈发重要的话题。作为神经网络解释性方法的一个分支,网络的路径可解释性受到了越来越多的关注。文中特别探讨了关键数据路由路径(Critical Data Routing Path,CDRP)这一面向网络路径的可解释方法。首先,通过 Score-CAM(Score-Class Activation Map)方法分析了 CDRP 在输入域上的路径可视化归因,指出 CDRP 方法在语义层面的潜在缺陷。然后,提出了一种语义引导的 Score-CDRP 方法,从方法机理上提升了 CDRP 与原始神经网络的语义一致性。最后,通过实验从路径热力图可视化以及相应的预测与定位精度等角度验证了 Score-CDRP 方法相较于 CDRP 的合理性、有效性和鲁棒性。

关键词:计算机视觉;深度神经网络;神经网络可解释性;特征可视化;网络剪枝;热力图
中图分类号　TP183

Semantic-guided Neural Network Critical Data Routing Path

ZHU Fukun[1],TENG Zhen[2],SHAO Wenze[1],GE Qi[1] and SUN Yubao[3]
1 School of Communications and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
2 Bell Honors School,Nanjing University of Posts and Telecommunications,Nanjing 210042,China
3 Engineering Research Center for Digital Forensics Ministry of Education,Nanjing University of Information Science and Technology,Nanjing 210044,China

Abstract　In recent years,with the popularity of artificial intelligence in various fields,it has become an increasingly important topic to study the interpretable methods of neural networks and understand their running principles. As a subfield of neural network interpretability methods,the interpretability of network pathways garners increasing attention. This paper particularly focuses on the critical data routing path(CDRP),an interpretable method for network pathways. Firstly,the routing path visualization attribution of CDRP in the input domain is analyzed by use of the score-class activation map(Score-CAM) method,pointing out the inherent defects of the CDRP approach in terms of semantics. Then a channel semantic guided CDRP method termed as Score-CDRP is proposed,which improves the semantic consistency between the original deep neural network and its corresponding CDRP from the perspective of method mechanism. Lastly,experimental results demonstrate that the proposed Score-CDRP approach is more reasonable,effective and robust than CDRP in terms of visualization of the routing path heatmap as well as its corresponding prediction and localization accuracy.

Keywords　Computer vision,Deep neural networks,Interpretability of neural networks,Feature visualization,Network pruning,Heatmap

1　引言

神经网络已被广泛应用于图像和语音识别、自然语言处理、医学诊断等领域,甚至在一些复杂任务中展现出了超越人类的水平[1-2]。然而,由于神经网络的高复杂性和不确定性,各种网络模型通常被视为黑盒[3],即使它们的准确率很高,人们也难以理解它们的决策过程。在某些情况下,理解网络的决策过程是必要的。例如,当神经网络用于医学诊断时,医生必须能够理解网络为何做出某个决策。此外,神经网络的不透明性可能导致模型的不公平性和错误决策。因此,研究神经网络的可解释性具有重要意义。理解了神经网络的决策过程,就可以更好地解释模型的行为、改进模型的性能、减少

到稿日期:2023-09-19　返修日期:2024-03-06
基金项目:国家自然科学基金(61771250,61972213)
This work was supported by the National Natural Science Foundation of China(61771250,61972213).
通信作者:邵文泽(shaowenze@njupt.edu.cn)

模型的不确定性和提高模型的适应性。

目前，已有多种方法被用于解释神经网络的运作机理和决策过程，如：深度可视化(Deep Visualization)[4]、激活最大化(Activation Maximization)[5]、类激活映射(Class Activation Map)[6]、相关性分数逐层传播(Layer-wise Relevance Propagation)[7]、导向反向传播(Guided Back Propagation)[8]等。

Wang 等[9]从关键数据路由路径(CDRP)的角度探讨了深度神经网络的解释性问题。通过引入面向神经元的控制门稀疏约束，CDRP 方法旨在实现对决策重要的关键神经元的稀疏鉴别。实验证明，CDRP 方法提取的决策路径在对抗样本检测和网络压缩剪枝方面均有着良好表现。

然而，通过 CDRP 的稀疏优化准则发现，CDRP 方法在保证路径决策结果与网络决策相同的条件下迭代选择路径节点，在解释机理上难以确保 CDRP 与原始神经网络所关注的语义信息高度一致。为此，本文开展了如下创新工作：

1)发现了深度神经网络的路径解释方法 CDRP 的语义归因缺陷。本文借助 Score-CAM 的可视化框架，分析了 CDRP 在输入图像上的语义归因，以路径热力图的可视化方式直观发现了 CDRP 的语义归因缺陷。

2) 提出了语义引导的 Score-CDRP。通过在 CDRP 的稀疏优化准则中新引入逐个神经元对于网络决策重要性的特征得分这一定量语义引导，实现了神经元控制门的自适应稀疏优化，从方法机理上提升了路由路径与原始神经网络的一致性，使路径更准确地代表网络决策。

3)定性定量验证。通过路径热力图可视化以及从相应的预测与定位精度等角度分析比较了 CDRP 和 Score-CDRP 在输入图像上的语义归因差异，验证了 Score-CDRP 方法相较于 CDRP 的合理性、有效性和鲁棒性。

2 相关工作

2.1 关键数据路由路径(CDRP)

为了更好地定量分析网络的每个组成部分对决策的影响，同时在 DNN 巨量的参数中选择对决策有重要贡献的稀疏网络节点，文献[9]提出了一种关键数据路由路径算法(CDRP)。CDRP 可识别每个给定输入的关键路径，并跟踪中间层的网络响应特性。

如图 1 所示，对于一个有 L 层输出通道的深度神经网络模型，CDRP 通过控制门 $\Lambda=[\lambda_1,\lambda_2,\cdots,\lambda_k,\lambda_{k+1},\cdots,\lambda_L]$ 表示路径，λ_k 表示第 k 层输出通道对应的控制门，网络每层的控制门乘该层对应的输出通道，得到实际的路由节点，连接起来的各层路由节点即为网络的路由路径。因此，将网络各层控制门与对应的网络输出通道相关联，即可得到网络的路径 v。为了得到稀疏且有意义的路由路径[10-11]，对控制门 Λ 的取值做出了如下规范：1)控制门必须取正值；2)控制门的值必须多数接近 0。根据以上规范，Wang 等[9]受知识蒸馏[12]的启发，设计了蒸馏导向路由(Distillation Guided Routing，DGR)算法。该算法通过对路径预测结果与网络预测结果之间的交叉熵进行稀疏正则优化获取 CDRP。

路径 v 可表示为：

$$v=concat([\lambda_1^*,\lambda_2^*,\cdots,\lambda_k^*,\cdots,\lambda_L^*]) \tag{1}$$

其中，λ_k^* 表示通过 CDRP 方法得到的第 k 层输出通道对应的控制门，$concat$ 表示将控制门依次连接的操作。

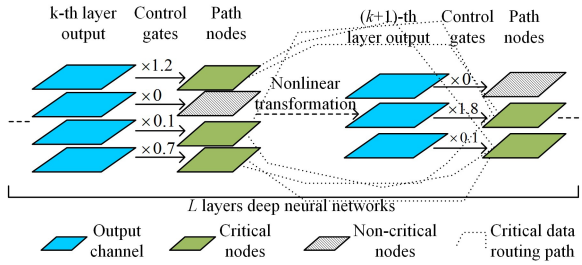


图 1 基于控制门的网络路由路径表示

Fig. 1 Routing path representation of control gate based network

2.2 类激活映射(CAM)

类激活映射(Class Activation Map，CAM)[6]通过热力图反映网络在输入域中所关注的区域，即对网络的决策进行归因。近年来，基于原始的 CAM 理念，衍生发展出了多种实用性更广、归因更加准确的 CAM。这些 CAM 可分为两种类别。1)基于梯度的 CAM，如：Grad-CAM[13]，Grad-CAM++[14]，Integrated Grad-CAM[15]，Layer-CAM[16]等。2)基于非梯度的 CAM，如：Score-CAM[17]，Augmented Score-CAM[18]等。其中，Score-CAM[17]由于可以避免产生梯度噪声，同时适用于多种网络结构，因此被广泛应用于网络的可视化解释。

此外，Kahkzar 等[19]还提出了路径梯度(Pathway Gradient)方法，该方法对剪枝之后的子网络进行梯度反向传播，得到输入域上对应的梯度，然后对输入域上的梯度进行可视化，进而得到网络路径的归因结果。本文将借助路径梯度这一方法，定性分析 CDRP 以不同稀疏度进行剪枝的子网络结果，并与 Score-CAM 方法得到的路径归因结果进行对照分析，以验证基于 Score-CAM 的路径归因分析的合理性。

3 提出方法

作为一种路径解释性方法，Wang 等[9]和 Kahzar 等[19]均证明了关键数据路由路径(CDRP)在对抗样本鉴别、路径稀疏性以及死亡神经元占比等方面[9,19-21]有良好的性能。然而，当前文献还缺乏对路径直观的语义归因分析。对于网络输入所对应的 CDRP，若能知晓路径在输入图像上的归因和定位结果，则有助于进一步分析与提升路径解释的合理性。

3.1 CDRP 的语义归因分析

为了对 CDRP 进行语义归因，本文进一步分析了 CDRP 的表示方法。由 2.1 节可知，路径节点通过控制门来表示：神经网络每层的输出通道在与控制门的数值相乘之后，得到网络实际的路径节点。因此，控制门在不改变输出通道维度的情况下，抑制了对网络的单次决策有着较小影响的通道，同时放大了对网络决策有重要影响的输出通道。如果将控制门改变后的输出通道用于网络的特征可视化，就可在一定程度上反映 CDRP 所关心的输入特征。同时，值得注意的是，Score-CAM[17]作为一种对输出通道进行加权的特征可视化方法，通过对输出通道进行 CIC(Channel-wise Increase of Confidence)计算，得到通道的重要性，并将该重要性作为通道权重。

考虑到 CDRP 和 Score-CAM 的特性，本文利用了 Score-CAM 的方法框架分析 CDRP 的特征归因：将控制门与网络原始的输出通道相乘，并将控制门改变后的输出通道用于 CIC 计算（如式(2)所示），得到控制门影响后的输出通道的贡献得分，并将贡献得分作为权重用于 CAM 的计算（如式(3)所示），最终得到网络路径在输入域中的归因。

$$\alpha_{k,l}^{c}=f_{\theta}(X\circ s(Up(A_k^l\cdot\lambda_k^l))) \tag{2}$$

$$L^{c}=ReLU(\sum_{l}\alpha_{k,l}^{c}(A_k^l\cdot\lambda_k^l)) \tag{3}$$

其中，A_k^l 表示网络第 k 层第 l 个输出通道，λ_k^l 表示 A_k^l 对应的控制门，$A_k^l\cdot\lambda_k^l$ 即为网络第 k 层第 l 个网络节点的输出，$f_\theta(\cdot)$表示含有控制门表示的整个路径的网络，X 表示网络的原始输入，$Up(\cdot)$表示对 A_k^l 上采样至输入维度的操作，$s(\cdot)$是将网络输入规范化到区间[0,1]的规范化函数，$\circ$ 表示 Hadamard 积的运算，$\alpha_{k,l}^{c}$即为第 k 层第 l 个路径节点面向目标类 c 的贡献得分。将贡献得分作为权重对路径节点输出进行加权，即可得到路径在目标类 c 上对应的类激活映射 L^c。

利用 Score-CAM 方法框架对 CDRP 路径归因进行可视化的具体流程如图 2 所示。在流程开始前，需要获取输入样本 CDRP 的控制门，并将控制门与网络中对应的输出通道相乘，得到含有样本对应 CDRP 的网络模型。流程主要包括 Phase1 和 Phase2 两个阶段。在 Phase1 阶段，将样本输入到包含样本对应 CDRP 的网络模型，获得网络最后一层卷积层的各路径节点的输出通道，之后将输出通道上采样至输入维度并进行归一化处理（对应式(2)中的 $s(Up(A_k^l\cdot\lambda_k^l))$，其中 k 为最后一层卷积层对应的层数）；在 Phase2 阶段，对处理后的节点输出和输入样本进行 Hadamard 乘积计算，并将 Hadamard 乘积运算的结果输入到包含样本 CDRP 的网络模型中，之后通过网络末层的 Softmax 运算得到网络最后一层卷积层各路径节点在输入样本目标类别 c 上的重要性得分（对应式(2)中的 $\alpha_{k,l}^{c}=f_\theta(X\circ s(Up(A_k^l\cdot\lambda_k^l)))$，其中 k 为最后一层卷积层对应的层数）。最后，将 Phase2 阶段获得的路径节点的重要性得分作为 Phase1 阶段获得的上采样及归一化处理后的路径输出通道进行加权求和，即可获取网络路径在输入域上的归因热力图（对应于式(3)）。

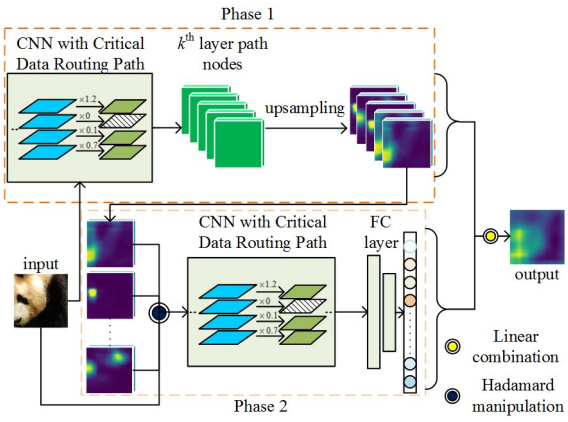
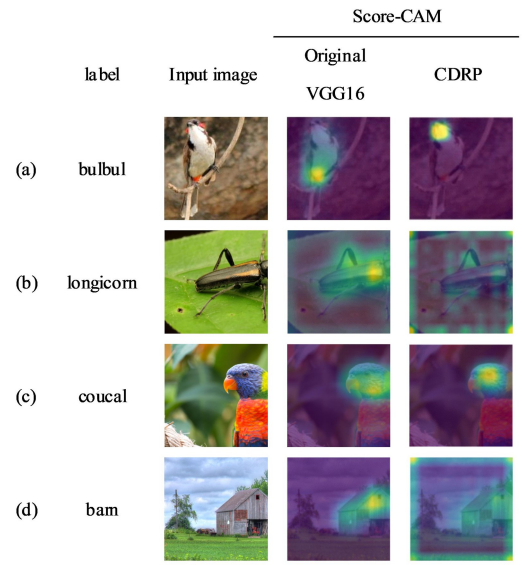


图 2 基于 Score-CAM 的 CDRP 路径热力图

Fig. 2 Heatmap of CDRP routing path based on Score-CAM



注：(a)，(c)路径归因合理，(b)，(d)路径归因错误。

图 3 CDRP 路径热力图示例

Fig. 3 Example of CDRP routing path heatmap

本文使用 VGG16[22]作为考查模型（VGG16 具有典型的卷积神经网络特征，同时近年来被广泛用于神经网络的可解释性研究，文献[7,9,17,19]均使用 VGG16 作为考查模型），在 ILSVRL2012 验证集[23]上随机选取了 1 000 张图片作为输入样本，采用文献[19]中的 CDRP 算法的超参设定获取了如图 3 所示的 CDRP 的路径热力图示例。其中，图 3(a)和图 3(c)路径归因合理，图 3(b)和图 3(d)路径归因错误。将 CDRP 的 Score-CAM 与原始网络（VGG16）的 Score-CAM 对照可以得出，如图 3(a)和图 3(c)所示，CDRP 在保持决策结果的前提下，往往只关注输入中那些决策所需的最重要特征区域。然而，如图 3(b)和图 3(d)所示，CDRP 也存在路径错误归因的情况：相对于原始网络对输入图像上对象的定位效果，CDRP 的归因定位更不准确。通过检查和分析得知，这些热力图归因失败的路径尽管保持了原有预测标签，但是路径的预测概率却显著地下降。原因之一在于，CDRP 方法在路径优化的过程中缺乏逐个神经元对网络决策重要性的语义引导，从而在解释机理上难以保证 CDRP 与原网络所关注信息的一致性。

3.2 Score-CDRP

根据 3.1 节的 CDRP 语义归因分析，本文在 CDRP 方法的优化目标函数中新引入基于 CIC 得分的语义信息，提出了一种语义引导的改进神经网络解释方法 Score-CDRP。

具体而言，一方面，新方法的优化目标函数保留了原始 CDRP 方法的交叉熵部分；另一方面，新方法对 CDRP 的控制门引入决策相关的语义约束，使得路径的稀疏优化过程受对应网络节点输出的语义信息的限制。考虑到 Score-CAM 中的通道得分的计算方法可以有效衡量输出通道对网络决策的贡献，而且计算方法适用于神经网络的不同卷积层，因此本文可为每层输出通道对应的控制门计算相应的 CIC 通道得分。在此基础上，路由路径的稀疏正则化可表示为各个输出通道的 CIC 得分与对应的控制门的线性加权组合。为此，新方法 Score-CDRP 的优化目标函数可表述为：

$$\min_{\Lambda} Loss(f_\theta(x), f_\theta(x;\Lambda)) + \gamma \sum_k \sum_l \alpha_{k,l}^c \cdot |\lambda_{k,l}| \quad \text{s.t. } \lambda_{k,l} \geqslant 0, k=0,1,2,\cdots,L \tag{4}$$

其中，$Loss(f_\theta(x), f_\theta(x;\Lambda))$表示原始神经网络 $f_\theta(\cdot)$与控制门指示的路由路径网络 $f_\theta(\cdot;\boldsymbol{\Lambda})$关于输入 x 的预测交叉熵，$\alpha_{k,l}^c$表示网络第 k 层第 l 个通道在目标类 c 上的 CIC 得分，$\lambda_{k,l}$表示网络第 k 层第 l 个控制门，γ 表示平衡参数。利用随机梯度下降(Stochastic Gradient Descent，SGD)算法，式(4)的优化目标最小化问题最终可由算法 1 具体实现。

算法 1 引入通道得分的蒸馏导向路由算法

输入：输入图像 x，预训练网络 $f_\theta(\cdot)$，控制门 $\boldsymbol{\Lambda}$，平衡参数 γ，最大迭代次数 T，SGD 优化器，网络的卷积集合 Layers

输出：语义引导的路由路径 v

1. 初始化控制门 $\boldsymbol{\Lambda}$ /* 将各节点对应控制门初始化为 1 */
2. 计算 x 在 $f_\theta(\cdot)$上的预测类别 c
3. for l in Layers do
4. 计算并保存第 l 层卷积层输出通道在目标类 c 上的 CIC 得分
5. end for
6. 计算初始损失 cur_loss，并保存为 minimum
7. for 1 to T do
8. 通过 SGD 优化更新控制门 $\boldsymbol{\Lambda}$，并保持对 $\boldsymbol{\Lambda}$ 的约束($\lambda_{k,l} \in [0,10]$)
9. 通过式(4)计算当前损失 cur_loss
10. 新的预测类别记录为 j
11. if c==j
12. if cur_loss $\leqslant$ minimum
13. 将控制门向量 $\boldsymbol{\Lambda}$ 保存到 v 中
14. end if
15. end if
16. end for

为了保证比较的公平性，在 Score-CDRP 和 CDRP 的优化算法超参数上均采用了与文献[19]一致的设定：优化器学习率 $lr=0.01$，迭代次数 $T=100$，优化器动量 $Momentum=0.9$，正则化项平衡参数 $\gamma=0.01$，权值衰减 $WeightDecay=0$。

在算法的时间复杂度方面，从获取路径所需的计算量这一角度来看，CDRP 和 Score-CDRP 方法都使用了 SGD 优化输入样本对应路径和网络预测的损失函数，两者的时间复杂度主要由 SGD 优化部分所决定，只在损失函数的计算上存在差异。假设 SGD 的迭代次数为 T，每次迭代更新路径节点参数需要的运算次数为 G。对于 CDRP 方法而言，计算 CDRP 优化目标所需运算可以分为两部分：第一部分计算路径预测和网络预测的交叉熵损失，所需的运算次数为 K_1；第二部分为计算 $L1$ 正则化项，所需运算次数为 K_2。因此对于 CDRP 方法，生成一个样本对应路径所需的计算次数为：$t_{\text{CDRP}}=T*(G+K_1+K_2)$。对于 Score-CDRP 方法，只在优化目标第二部分计算 CIC 得分加权的正则化项上所需的运算次数相对于 CDRP 方法有所变化：Score-CDRP 通过加权求和计算正则化项所需的运算次数为 $2*K_2$。因此 Score-CDRP 方法生成一个样本对应路径所需的计算次数：$t_{\text{Score-CDRP}}=T*(G+K_1+2*K_2)$。尽管 Score-CDRP 需要事先计算网络各卷积层的 CIC 得分，但相对于 SGD 迭代部分的计算量，CIC 得分所需计算量较小，对算法整体性能影响不大。通过大 O 表示法表示两者的时间复杂度，则 CDRP 方法的时间复杂度为 $T_{\text{CDRP}}=O(T*n)$，Score-CDRP 方法的时间复杂度为 $T_{\text{Score-CDRP}}=O(T*n)$，其中 T 表示迭代次数，n 表示每次迭代的运算次数，两者的时间复杂度同为平方阶复杂度。在算法的空间复杂度方面，由于 CDRP 与 Score-CDRP 方法同为事后可解释性方法，其空间复杂度主要受到目标模型参数量的影响，因此两者空间复杂度基本一致。

类似 3.1 节，将 VGG16 作为考查模型，图 4 给出了 Score-CDRP 输出的路径热力图示例。如图 4(b)－图 4(d)所示，Score-CDRP 的路由路径近乎保持了与原始网络的语义归因一致性；相较于 CDRP 方法的低精度或错误性语义归因，Score-CDRP 的路径归因更合理、更精确。

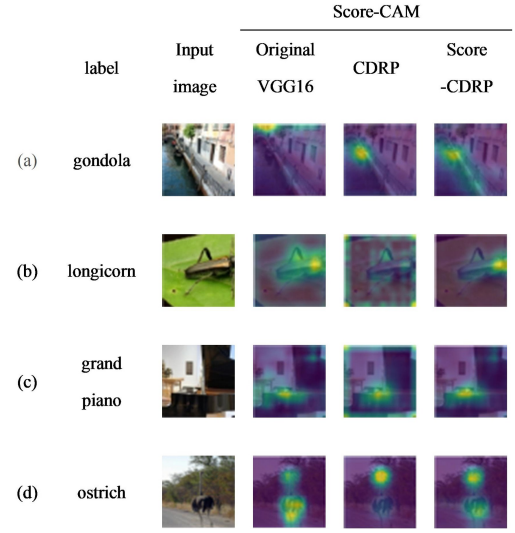


图 4 CDRP 与 Score-CDRP 的路径热力图示例

Fig. 4 Routing path heatmap of CDRP and Score-CDRP

此外，从图 4(a)中观察到一个有趣的现象：原始网络实现了正确的标签预测，即 gondola(威尼斯小划船)，但是相应的归因热力图却聚焦于左上角的建筑区域；而 CDRP 和 Score-CDRP 均实现了正确归因。特别地，对应 Score-CDRP 的网络预测过程不仅将注意力聚焦到了 gondola 的船身，而且兼顾了 gondola 所在的河岸。值得指出的是，Score-CDRP 的合理性、精确性和鲁棒性只是本文简单有效地引入通道语义而实现的。第 4 章将进一步量化评估 CDRP 和 Score-CDRP 的路径可解释性。

4 实验结果与分析

本章将通过实验检验路径语义归因是否合理以及 Score-CDRP 相较于原始 CDRP 方法在有效性和鲁棒性方面是否有所提升。由于目前暂时没有公开其他的基于 CDRP 改进的面向网络路径的解释方法，本章实验只比较了 Score-CDRP 和原始 CDRP 方法的性能，包括针对路径的定性分析实验以及量化评估实验。实验中所有路径都是以 ImageNet 数据集[23]上预训练的 VGG16 网络模型[22]为考查模型。所有实验均在 GTX1080ti 上运行，实验所需软件环境为 pyTorch1.10.0(GPU 版本)，生成 Score-CDRP 的优化算法通过 SGD 优化器实现，Score-CDRP 和 CDRP 方法所需超参数与 3.2 节的超参设置一致。

4.1 定性分析实验

如图 3 和图 4 所示，本文主要借助基于 Score-CAM 的路径的归因可视化，实现网络的路径解释性方法的定性、定量评估。在这之前，首先对这种评估方式的合理性和有效性做对照性验证，以保证路径的 Score-CAM 确实对路径所“关心”的特征区域进行了正确的归因。具体地，本文借助文献[19]中进行的路径剪枝实验以及所提出的路径梯度方法，从另一个角度对路径进行可视化分析。对于相同的路径，若路径梯度可视化与路径的 Score-CAM 可视化都“关心”了相似的特征区域，则认为完成了上文所述的对照性验证。为此，在 ILSVRL2012[23] 验证集随机选取图片，分别将 CDRP 和 Score-CDRP 两种方法的输出通道对应的控制门数值作为网络的剪枝准则，可得到不同剪枝率下的路径梯度归因图。图 5 给出了 CDRP 和 Score-CDRP 的路径热力图和路径梯度归因图的可视化对照。由结果可知，对于本文提出的 Score-CDRP，不管是路径热力图还是根据路径剪枝得到的路径梯度归因图，均在输入图像上进行了较相似的归因。如图 5 中的熊猫，Score-CDRP 的路径热力图和不同剪枝率下的路径梯度归因图，均成功归因到了熊猫的眼部和嘴部区域。

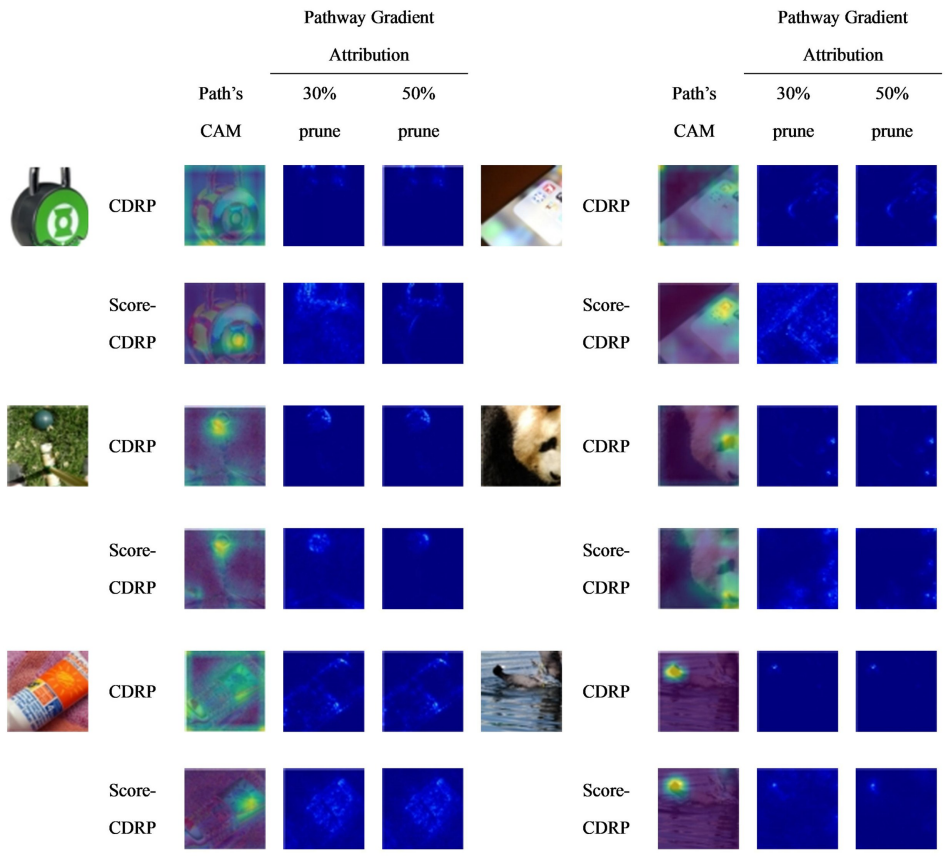


图 5 CDRP 和 Score-CDRP 的路径热力图、路径梯度归因图示例

Fig. 5 Routing path heatmap and pathway gradient of CDRP and Score-CDRP

值得注意的是图 5 中的闹钟示例：路径热力图“关心”的是小闹钟区域，而路径梯度归因图本质上“关心”的则是大闹钟区域。在此示例中，路径热力图一定程度上是路径梯度归因图的子集。此外，图 5 的对照实验再次验证了 Score-CDRP 相较于 CDRP 的优越性。例如，在闹钟示例中，CDRP 的路径热力图是不合理的甚至是错误的，而 CDRP 的路径梯度归因图则是完全错误的。

此外，为了对比两种路径的 Score-CAM 在多个同类目标图像中的定位性能，本文使用 ILSVRL2012[23] 中的多目标图像进行了多目标定位实验：比较了原始 VGG16 网络、CDRP 以及 Score-CDRP 的热力图在同类多目标图像上的定位效果。部分实验结果如图 6 所示，相对于 CDRP，Score-CDRP 所代表的路径可以更均衡地定位到图像中的多个目标，更接近原始网络所关注的信息。

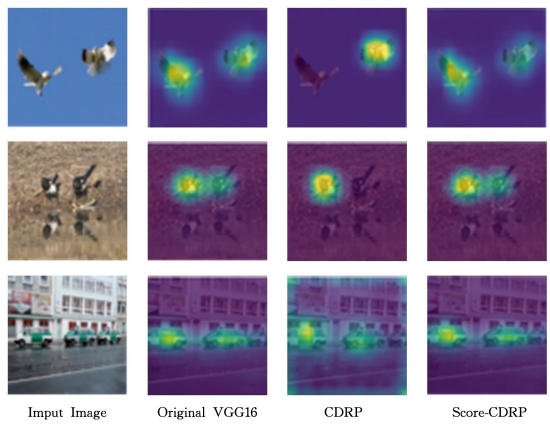


图 6 多目标样本的 CDRP 与 Score-CDRP 热力图可视化结果

Fig. 6 Visualization results of CDRP and Score-CDRP heatmaps of multi-target samples

4.2 量化评估实验

在定量评估 Score-CDRP 方法和原始 CDRP 方法的性能表现方面，首先统计了相应路径网络的预测置信度，之后借助上文的路径热力图归因方法，利用热力图的评价指标完成两种路径解释性方法的定量评估。

由于控制门表示的路径对输入信息的传播起到抑制或加强的作用，因此路径网络对输入的预测置信度可直观反映路径解释的可信度。为此，选取 ImageNet-compatible 数据集作为实验用例(该数据集包含 1 000 张图片，曾用于 NIPS2017 对抗赛)，考查模型依然为 VGG16。本文计算了 CDRP 和 Score-CDRP 在 ImageNet-compatible 数据集上的所有样本的平均预测置信度，两种路径的平均预测置信度如表 1 所列。由结果可知，相较于原始 CDRP 方法，Score-CDRP 所代表的路径网络对于输入有着更高的平均置信度，一定程度上验证了 Score-CDRP 实现了可信度更高的路径解释。

表 1 CDRP 和 Score-CDRP 的平均预测置信度

Table 1 Average prediction confidence of CDRP and Score-CDRP

(%)

Model/Path	Average prediction confidence
Original VGG16	71.59
CDRP	67.65
Score-CDRP	68.41

进一步，本文采用热力图的置信增强(Increase Of Confidence，IOC)和平均下降(Average Drop，AD)指标[14]来评估对路径解释的可信度。置信增强和平均下降指标如式(5)和式(6)所示：

$$IOC=\sum_{i=1}^{N}(\mathrm{Sign}(Y_i^c<O_i^c)/N) \tag{5}$$

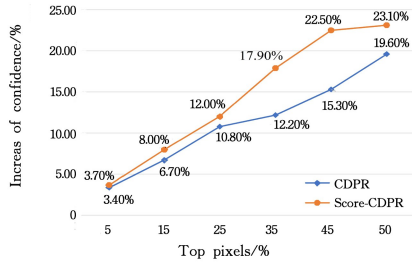
$$AD=\sum_{i=1}^{N}(\max(0,Y_i^c-O_i^c)/Y_i^c) \tag{6}$$

其中，Y_i^c 是图片 i 在类别 c 上的预测得分，O_i^c 是将路径热力图在输入图像的解释映射区域作为输入在类别 c 上的预测得分。Sign(·)为一个指示器函数，当输入为真时，返回 1。上述两个指标用于评估热力图中的解释性区域对于深度网络判断的重要性，也即网络决策是否由热力图的解释性区域所主导。本文按不同比例从热力图选取像素作为解释性区域[24]。以 VGG16 的最后一个卷积层的路径热力图为例，图 7 给出了 CDRP 和 Score-CDRP 的置信增强与平均下降结果。

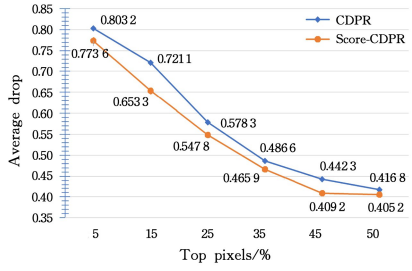
由结果可知，当选择不同的像素比例作为解释性区域时，Score-CDRP 的平均下降指标总要小于 CDRP(如图 7(a)所示)，同时 Score-CDRP 的置信增强指标总是强于 CDRP(如图 7(b)所示)。因而，与 CDRP 相比，Score-CDRP 的路径热力图聚焦的特征区域对于标签预测更为精准，因此 Score-CDRP 对于路径解释的可信度更高。

最后，通过定量评估路径热力图的解释性区域的定位准确性来验证 Score-CDRP 的合理性。具体地，本文采用基于能量的指向比赛(Proportion)这一指标[16]对解释性区域与目标区域的契合程度进行评估。基于能量的指向比赛的计算式如式(7)所示。

$$Proportion=\left(\frac{\sum L^c_{(i,j)\in bbox}}{\sum L^c_{(i,j)\in bbox}+\sum L^c_{(i,j)\notin bbox}}\right) \tag{7}$$



(a)置信增强



(b)平均下降

图 7 路径 CAM 的定量评估结果

Fig. 7 Quantitative evaluation results of routing path CAM

首先用目标类别的边界框对输入图像进行二值化，并将内部区域分配为 1，外部区域分配为 0，之后将其与路径热力图进行点乘并求和，得到边界框中的能量。本文从 ILSVRC2012 验证集随机选取 500 张带有边框信息的图片[23]，比较了原始 VGG16 的热力图以及 CDRP 和 Score-CDRP 的路径热力图在实验数据集上的表现，结果如表 2 所列。由实验结果可知，相较于 CDRP，Score-CDRP 聚焦的解释性区域与目标区域有着更好的契合程度，路径归因更为精准，从而再次验证了 Score-CDRP 可信度更高的路径解释能力。

表 2 基于能量的指向比赛结果

Table 2 Results of energy-based pointing game

Model/Path	*Proportion*
Original VGG16	0.4054
CDRP	0.3802
Score-CDRP	0.3956

结束语 本文探讨了 CDRP 这一面向深度神经网络的路径可解释性方法。通过 Score-CAM 方法实现了 CDRP 在输入域上的路径可视化归因，指出了 CDRP 方法在语义层面的潜在缺陷。以此为动机，基于 CIC 得分进一步提出了一种语义引导的 Score-CDRP 方法，从模型机理上提升了 CDRP 与原始神经网络的语义一致性。最后，实验结果从路径热力图的可视化及相应的预测与定位精度等方面验证了 Score-CDRP 算法相较于 CDRP 算法的合理性、有效性和鲁棒性。

虽然 Score-CDRP 相对于 CDRP 算法在多个方面都有所改善，但仍有进一步改进的空间：1)在对较大的网络模型使用引入 Score-CDRP 的蒸馏导向算法会相对耗时，未来将降低算法的时间复杂度；2)根据本文的定量实验结果可以发现，Score-CDRP 在路径归因准确性方面还有可以提升的空间。因此如何改善 Score-CDRP 的性能，也是未来需要讨论的一个问题。

参 考 文 献

[1] SUBAKAN C，RAVANELLI M，CORNELL S，et al. Attention is all you need in speech separation[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing. (ICASSP 2021)IEEE，2021：21-25.

[2] ZHU X，LYU S，WANG X，et al. TPH-YOLOv5：Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：2778-2788.

[3] RICCARDO G，ANNA M，SALVATORE R，et al. A Survey Of Methods For Explaining Black Box Models[J]. ACM Computing Surveys，2018，51(5)：1-42.

[4] YOSINSKI J，CLUNE J，NGUYEN A，et al. Understanding neural networks through deep visualization[J]. arXiv：1506.06579，2015.

[5] ERHAN D，BENGIO Y，COURVILLE A，et al. Visualizing higher-layer features of a deep network[J]. University of Montreal，2009，1341(3)：1-13.

[6] ZHOU B，KHOSLA A，LAPEDRIZA A，et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：2921-2929.

[7] BACH S，BINDER A，MONTAVON G，et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PloS One，2015，10(7)：e0130140.

[8] SPRINGENBERG J T，DOSOVITSKIY A，BROX T，et al. Striving for simplicity：The all convolutional net[J]. arXiv：1412.6806，2014.

[9] WANG Y，SU H，ZHANG B，et al. Interpret neural networks by identifying critical data routing paths[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：8906-8914.

[10] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society：Series B(Methodological)，1996，58(1)：267-288.

[11] HOEFLER T，ALISTARH D，BEN-NUN T，et al. Sparsity in deep learning：Pruning and growth for efficient inference and training in neural networks[J]. The Journal of Machine Learning Research，2021，22(1)：10882-11005.

[12] HINTON G，VINYALS O，DEAN J. Distilling the Knowledge in a Neural Network[J]. Computer Science，2015，14(7)：38-39.

[13] SELVARAJU R R，COGSWELL M，DAS A，et al. Grad-cam：Visual explanations from deep networks via gradient-based localization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：618-626.

[14] CHATTOPADHAY A，SARKAR A，HOWLADER P，et al. Grad-cam＋＋：Generalized gradient-based visual explanations for deep convolutional networks[C]// 2018 IEEE Winter Conference on Applications of Computer Vision. IEEE，2018：839-847.

[15] SATTARZADEH S，SUDHAKAR M，PLATANIOTIS K N，et al. Integrated grad-cam：Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring[C]// ICASSP 2021-2021 IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2021：1775-1779.

[16] JIANG P，ZHANG C，HOU Q，et al. LayerCAM：Exploring Hierarchical Class Activation Maps for Localization. [J]. IEEE Transactions on Image Processing：a Publication of the IEEE Signal Processing Society，2021，30：5875-5888.

[17] WANG H，WANG Z，DU M，et al. Score-CAM：Score-weighted visual explanations for convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020：24-25.

[18] IBRAHIM R，SHAFIQ M O. Augmented Score-CAM：High resolution visual interpretations for deep neural networks [J]. Knowledge-Based Systems，2022，252：109287.

[19] KHAKZAR A，BASELIZADEH S，KHANDUJA S，et al. Neural response interpretation through the lens of critical pathways [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：13528-13538.

[20] ZHANG Y，TIŇO P，LEONARDIS A，et al. A survey on neural network interpretability[J]. IEEE Transactions on Emerging Topics in Computational Intelligence，2021，5(5)：726-742.

[21] QIU Y，LENG J，GUO C，et al. Adversarial defense through network profiling based path extraction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：4777-4786.

[22] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv：1409.1556，2014.

[23] RUSSAKOVSKY O，DENG J，SU H，et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision，2015，115：211-252.

[24] DESAI S，RAMASWAMY H G. Ablation-CAM：Visual Explanations for Deep Convolutional Network via Gradient-free Localization[C]// Workshop on Applications of Computer Vision. IEEE，2020：972-980.

ZHU Fukun，born in 1999，postgraduate. His main research interests include interpretability and adversarial transferability of deep learning models.

SHAO Wenze，born in 1981，Ph.D，professor. His main research interests include computational imaging，computer vision and machine learning.

（责任编辑：柯颖）