计算机科学
COMPUTER SCIENCE

基于局部注意力图互迁移的可解释性优化方法

成科扬，王宁，崔宏纲，詹永照

引用本文

成科扬，王宁，崔宏纲，詹永照. 基于局部注意力图互迁移的可解释性优化方法[J]. 计算机科学, 2022, 49(5): 64-70.

CHENG Ke-yang, WANG Ning, CUI Hong-gang, ZHAN Yong-zhao. Interpretability Optimization Method Based on Mutual Transfer of Local Attention Map[J]. Computer Science, 2022, 49(5): 64-70.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400176

基于局部注意力图互迁移的可解释性优化方法

成科扬　王　宁　崔宏纲　詹永照
江苏大学计算机科学与通信工程学院　江苏 镇江 212013

摘　要　目前,深度学习模型已被广泛部署于各个工业领域。然而,深度学习模型具有的复杂性与不可解释性已成为其应用于高风险领域最主要的瓶颈。在深度学习模型可解释性方法中,最重要的方法是可视化解释方法,其中注意力图是可视化解释方法的主要表现方式,可通过对样本图像中的决策区域进行标注,来直观地展示模型决策依据。目前已有的基于注意力图的可视化解释方法中,单一模型注意力图存在标注区域易出现标注错误而造成可视化可解释性置信度不足的问题。针对上述问题,文中提出了一种基于局部注意力图互迁移的可解释性优化方法,用于提升模型注意力图的标注准确度,展示出精准的决策区域,加强视觉层面对模型决策依据的可解释性。具体表现为:采用轻量模型构建互迁移网络结构,于单一模型层间提取特征图并进行叠加,对全局注意力图进行局部划分,使用皮尔逊相关系数对模型间对应的局部注意力图进行相似度度量,随后将局部注意力图进行正则化并结合交叉熵函数对模型注意力图进行迁移。实验结果表明,所提算法显著提升了模型注意力图标注的准确性,并分别实现了28.2%的平均下降率和29.5%的平均增长率,与最先进的算法相比,其在平均下降率方面实现了3.3%的提升。实验结果表明,所提算法能成功地找出样本图像中预测标签最相关区域,而不局限于视觉可视化区域;与现有的同类方法相比,所提方法能更准确地揭示原始CNN模型的决策依据。

关键词:可解释性;注意力图;区域划分;相似度;互迁移
中图法分类号　TP391.4

Interpretability Optimization Method Based on Mutual Transfer of Local Attention Map

CHENG Ke-yang,WANG Ning,CUI Hong-gang and ZHAN Yong-zhao
School of Computer Science and Communication Engineering,Jiangsu University,Zhenjiang,Jiangsu 212013,China

Abstract　At present,deep learning models have been widely deployed in various industrial fields. However,the complexity and inexplicability of deep learning model have become the main bottleneck of its application in high-risk fields. The most important method is the visual interpretation,in which the attention map is the main representation of the visual interpretation method. The decision area in the sample image can be marked to visually display the decision basis of the model. In the existing visual interpretation methods based on attention map,the single model attention map has the problem of insufficient confidence of visualization interpretability due to the annotation error easily appearing in the annotated region. To solve the above problems,this paper proposes an interpretable optimization method based on the mutual transfer of local attention map,aiming at improving the annotation accuracy of the model attention map and displaying the precise decision area,so as to strengthen the visual interpretable ability for the model decision basis. Specifically,the structure of the intermigration network is constructed by using the lightweight model,the feature maps are extracted and superimposed between the layers of the single model,and the global attention map is divided locally. Pearson correlation coefficient is used to measure the similarity of the corresponding local attention map between the models,and then the local attention map is regularized and transferred combined with the cross-entropy function.. Experimental results show that the proposed algorithm significantly improves the accuracy of the model attention map label accuracy. The proposed algorithm achieves an average drop rate of 28.2% and an average increase rate of 29.5%,respectively,and achieves an increase of 3.3% in the average decline rate compared with the most advanced algorithm. The above experiments show that the proposed algorithm can successfully find out the most responsive region in the sample image,rather than being limited to the visual visualization region. Compared with the existing similar methods,the proposed method can more accurately reveal the decision basis of the original CNN model.

Keywords　Interpretability,Attention diagram,Regional division,Similarity,Mutual transfer

到稿日期:2021-04-18　返修日期:2021-09-07
基金项目:国家自然科学基金(61972183,61672268);社会安全风险感知与防控大数据应用国家工程实验室主任基金项目
This work was supported by the National Natural Science Foundation of China(61972183,61672268) and Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data.
通信作者:成科扬(kycheng@ujs.edu.cn)

1 引言

目前，深度学习模型已被广泛部署于各个工业领域。然而，深度学习模型所具有的复杂性与不可解释性已成为其应用于高风险领域最主要的瓶颈。在深度学习模型可解释性方法中，最重要的方法为可视化解释方法，其中注意力图是可视化解释方法的主要表现方式，可通过对样本图像中的决策区域进行标注，来直观地展示模型的决策依据[1]。

目前利用现有的注意力图对模型进行可解释性的研究工作存在诸多缺陷，其主要问题在于注意力图标注区域出现偏差，不能准确地标注出模型做出决策的区域，导致模型的可解释性差，难以提高模型的可信度等问题。如图 1 所示，覆盖多个目标时，从视觉上难以获得模型精准的决策依据。

图 1 如何给出分类为苹果的依据？

Fig. 1 How to give the explanation of classification as apple?

本文研究的主要任务是解决目标模型构建的注意力图存在标注区域不准确而导致的可解释性较差的问题。本文的主要贡献为：

(1)提出了基于注意力图的互学习方法。该方法着眼于提升模型之间的可解释性。本文使用具备先验知识的模型，与另一种未经训练的模型相互学习，提升模型间的可解释性。

(2)提出了局部迁移注意力图的迁移方法，避免全局迁移对模型带来负迁移的影响。当注意力图部分区域相似度较低时，全局迁移会导致注意力图标注错误，使模型可解释性下降。本文将注意力图进行局部迁移，对模型注意力图进行区域划分，度量局部区域相似度并进行迁移，提升模型分类性能及注意力图标注准确性，从而有效提高模型的可解释性。

本文第 2 节对相关工作以及与后续章节密切相关的理论进行了详细介绍；第 3 节提出了具有可解释性的局部迁移互学习方法；第 4 节对实验结果进行分析，验证算法的有效性；最后总结全文。

2 相关工作

2.1 模型可视化方面

在卷积神经模型中，特征可视化具备的可解释性对直观理解模型决策具有重要作用。Yadav 等[2]提出了基于梯度的此类方法，该方法基于深度卷积模型中最大值分类的输出得分，对该类的特征进行了可视化。Zeiler 等[3]提出了反卷积模型，该模型与原始模型共享权重，并将某些特征投影到图像平面上，从而对图片进行可视化。通过引入正则化进一步优化激活最大化的方法，使可视化图像更清晰、更具可解释性。Gaur 等[4]提出了向上卷积模型，将 CNN 特征图反转为图像，将上卷积模型视为一种间接说明与特征图相对应的图像外观的工具。然而，与基于梯度的方法相比，上卷积方法在数学上无法确保可视化结果能准确反映 CNN 中的特征表示。

注意力可视化模型是通过可视化来解释深度学习模型表示的一种典型技术。基于注意力的机制最早是由 Larochelle 等[5]通过受限布尔兹曼机完成的。Selvaraju 等[6]提出 Grad-CAM，这是一种结合了引导反向传播和 CAM 的方法，此方法对分类结果贡献度较高的区域进行可视化，对模型预测提供了更好的视觉可解释性。与此同时，注意力图作为特征可视化的另一种重要方法，备受关注[7]。近年来，注意力机制开始应用于基于递归神经模型的机器翻译等方面及一些与 NLP 相关的任务[8]，同时也被用于计算机视觉领域。

2.2 模型迁移学习方面

在模型结构优化中，Hinton 等[9]提出了蒸馏模型，该模型利用预训练模型作为教师为学生模型提供额外的知识。实验结果表明，小模型通过模仿大模型估计的类别概率，表现出与大模型相近的性能。然而，蒸馏模型需要预训练教师模型作为先验知识[10]，且仅对小模型进行单向的知识传递，小模型在学习过程中难以向教师模型反馈信息，无法对训练过程进行优化，因此模型的准确性与可解释性具有局限性[11]。

在传统教师-学生模型间进行注意力图传递的过程中，模型能够学习到教师模型学习的注意力图与知识，从而取得很好的检测效果。在之前的工作中，常见的方法是模型进行全局的互相迁移学习，即选取整幅图像进行迁移。传统方法没有考虑到模型自身注意力图权重分布不同对迁移学习产生的影响。在大部分情况下，经过训练的模型的先验知识受到了教师模型的限制，造成其注意力图存在标注不准确甚至标注错误的问题。并且，当模型需要学习新的知识时，重复训练教师模型会导致开销变大。针对上述两个问题，本文提出了基于注意力图的互学习模型，在两种轻型模型之间进行注意力图的迁移学习，可以提升目标模型注意力图的准确性以及模型的可解释性，并且可以在学习新知识的同时降低开销。

基于上述事实，本文对注意力图迁移进行研究，在迁移过程中使用不同结构的模型同时进行训练，以互相学习，模型学习到不同的权重知识有助于提升模型的泛化能力以及模型的可解释性。

3 可解释性局部互迁移方法

针对目前模型特征图可视化方案中特征图标注区域不准确、标注错误导致模型可解释性下降的问题，本文提出了一种基于局部注意力图互迁移的可解释性优化方法。其主要包括注意力图构建算法、注意力图相似度度量算法及注意力图局部迁移算法 3 部分。该算法的结构流程图如图 2 所示。

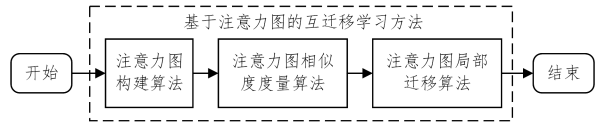

图 2 基于注意力图的可解释局部迁移互学习算法结构图

Fig. 2 Structure diagram of explanable local transfer mutual learning algorithm based on attention map

3.1 模型注意力图构建方法

本文提出了一种基于注意力图的可解释性局部迁移互

学习方法，提升了模型的可解释性。在迁移过程中，其主要使用两个轻量模型同时训练，使不同模型学习到不同的权重知识并于层间提取出特征图，从而构建注意力图。将注意力图进行划分后，对局部注意力图进行相似度度量及迁移学习。提升模型的泛化能力，使注意力图标注更为准确，从而提升模型的可解释性。模型的总体结构如图 3 所示。

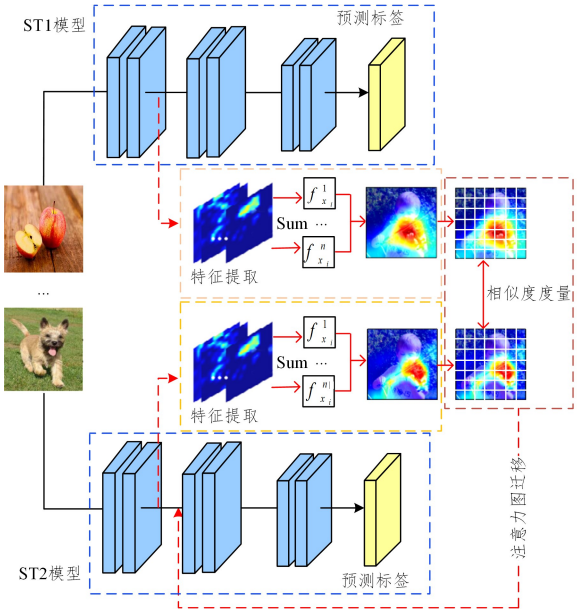

图 3 模型结构的示意图

Fig. 3 Schematic diagram of model structure

首先，在特征提取方面，以 ResNet-50 残差网络为例，ResNet-50 由若干个残差块组成，在每一个残差块后提取出模型特征图组，并根据卷积核通道数确定特征图数量，较低层级为低级边角特征，较高层级为全局特征。

在注意力图构建过程中，首先将上述训练过程中的模型特征图，即对应的 3D 激活张量 $\boldsymbol{A}\in f^{C\times H\times W}$ 作为输入，H 和 W 分别为特征图高度与宽度，C 为通道数。将 A 传入映射函数 F 后，输出空间注意图，如图 4 所示。

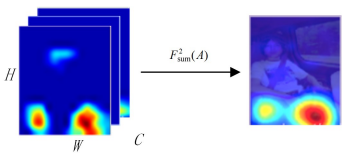

图 4 注意力图构建示意图

Fig. 4 Schematic diagram of attention map construction

上述映射函数 F 如式(1)所示：

$$F^2_{\mathrm{sum}}(\boldsymbol{A})=\sum_{i=1}^{C}|A_i|^2 \tag{1}$$

然后对注意力图进行相似度度量，选取模型构建的注意力图中高相似度的区域，并通过注意力图互迁移损失函数进行迁移。

3.2 注意力图相似度度量

为此，文中结合上述相互学习方法的优势，提出了模型层间注意图度量算法。由于不同模型层间对同一样本的注意力图权重是不同的，因此在相互学习网络中，学生网络需要对模型间的注意力图进行度量，从每个训练实例中找出并匹配相似特征，从而增加每个学生网络的后验熵，提升模型的准确度与泛化能力。因此，图像度量方式的选取至关重要。

首先，在 3.1 节中可获得基于模型间特征图构建的注意力图，调整模型输出特征映射的尺寸，假设模型层间注意力图尺寸为 $H\times W$。为保证上采样对注意力图空间信息的完整性，本节以特征图所在卷积层中卷积核的大小为基准，对注意力图进行划分。$K=\{1,2,\cdots,i,\cdots,k\}$ 为局部注意力图标号，局部注意力图如式(2)与式(3)所示：

$$Q_{S1}=\{Q^1_{S1},Q^2_{S1},\cdots,Q^i_{S1},\cdots,Q^k_{S1}\} \tag{2}$$

$$Q_{S2}=\{Q^1_{S2},Q^2_{S2},\cdots,Q^i_{S2},\cdots,Q^k_{S2}\} \tag{3}$$

其中，Q_{S1} 表示 ST1 模型局部注意力图组，Q_{S2} 表示 ST2 模型局部注意力图组，Q^i_{S1} 与 Q^i_{S2} 分别为对应局部注意力图 F^i_{S1} 与 F^i_{S2} 的向量表示形式，i 表示对应局部注意力图索引序号，k 为局部特征图总体数量。

在度量对应局部注意力图相似度方面，本文采用皮尔逊度量算法来度量注意力图间的相似度。其中，X 和 Y 表示两组变量，$\overline{X}$ 和 $\overline{Y}$ 表示两个变量的平均值，分别指 ST1 与 ST2 的局部注意力图。对向量进行归一化处理后，两个特征向量之间的距离如式(4)所示：

$$P_{x,y}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{4}$$

对于得到的局部注意力，通过皮尔逊距离进行相似度度量，得出注意力图向量间的距离 $P_{(Q^i_{S1},Q^i_{S2})}$。为了避免负迁移对注意力图标注带来的负面影响，通过设定阈值 λ 对相似度较高的区域进行迁移，对相似度较低的区域则丢弃。

由此，通过皮尔逊度量对应注意力图的相似度以及设定的阈值来确定 ST1 模型与 ST2 模型之间需要进行迁移的局部注意力图，以提升模型注意力图标注的准确性。本节提出的算法有效避免了负迁移对模型分类精度和注意力图标注的影响。

3.3 局部注意力迁移损失函数

局部注意力图将通过注意力迁移损失函数来实现 ST1 模型与 ST2 模型之间的迁移。令 S_1，S_2 和 W 分别表示 ST1 模型与 ST2 模型及各自的权重。所提损失函数如式(5)所示：

$$L_{AT}=L(W,x)+\frac{\beta}{2}\sum_{j\in I}\sum_{i=1}\left\|\frac{Q^{j(ki)}_{S_1}}{\|Q^{j(ki)}_{S_1}\|_2}-\frac{Q^{j(ki')}_{S_2}}{\|Q^{j(ki')}_{S_2}\|_2}\right\|_2 \tag{5}$$

损失函数由两部分组成，第一部分 $L(W,x)$ 为传统的标准交叉熵损失函数，表示模型在训练过程中对自身权重进行更新；第二部分由局部注意力图互学习构成，j 表示第 j 对模型层间注意力图的索引对，i 力局部注意力图索引对的序号。同时，使用二范数对局部注意力图进行正则化，即使用 $\frac{Q}{\|Q\|_2}$ 替换矢量化注意力图 Q。针对层间局部注意力图，通过 $\left\|\frac{Q^{j(i)}_{S_1}}{\|Q^{j(i)}_{S_1}\|_2}-\frac{Q^{j(i)}_{S_2}}{\|Q^{j(i)}_{S_2}\|_2}\right\|_2$ 使用欧氏距离衡量向量距离，得到损失函数并对模型进行训练。$\frac{\beta}{2}$ 为人工设定的权值，表示外部知识对网络模型的影响程度。

局部互迁移算法的伪代码如算法 1 所示。

算法 1 局部互迁移算法

输入：当前样本图像

输出：优化后的注意力图

```
1. 当前样本图像输入模型
2.   提取特征图组
3.   根据式(1)、式(2)得到相应全局注意力图
4.   将全局注意力图划分为 K 个部分，由式(2)、式(3)与式(2)、
     式(4)表示
5.   If i<K.length
6.     if 相似度大于 λ：
7.       保留对应区域 k(i)
8.       i++
9.     else
10.      丢弃对应区域 k(i)
11.      i++
12.  结束循环
13.  使用式(2)、式(5)对局部注意力图进行迁移
14. End for
```

基于注意力图的互学习方法适用于各种网络架构，以及由不同大小的混合网络组成的异构群组。具备先验知识的 ST1 模型向 ST2 模型迁移自身的注意力图，ST2 模型通过学习得到的特征权重对模型自身权重进行优化，在之后的注意力迁移损失函数中向 ST1 模型反馈自身注意力图，以辅助提升 ST1 模型的准确度。对 ST1 模型的注意力图进行改善，使注意力图标注区域更精准，从而进一步精确标注 ST1 模型的注意力图，在具有初步可解释性的注意力图的基础上，提升模型的可解释性。

4 实验结果与分析

4.1 实验配置

本文的实验环境为 TITAN-RTX，3.60 GHz 主频，32 GB 内存，500 GB 硬盘，一台 Windows 10 操作系统的主机和 110 GB 内存，7 TB 磁盘，NVIDIA tesla P100 显卡 16 GB×2，一台 Linux 操作系统的服务器，实验平台的软件配置是 PyCharm2016＋Python3.6＋OpenCV3。

数据集采用 CIFAR-10 与 CIFAR-100 数据集。CIFAR-10 数据集包含 60 000 张彩色图像，图像尺寸为 32 * 32，分为 10 个类，每类 6 000 张图，其中 50 000 张用于训练，构成 5 个训练批，每一批 10 000 张图；另外 10 000 用于测试，单独构成一批。测试批的数据取自 10 类中的每一类，每一类随机取 1 000 张图像。将剩下的随机排列组成训练批。一个训练批中的各类图像数量并不一定相同。总的来看，训练批每一类都有 5 000 张图。CIFAR-100 包含 100 个类，每个类包含 600 张图像，每类各有 500 张训练图像和 100 张测试图像。

4.2 评价标准

可视化可解释性因为其最终目的的特殊性以及影响因素的复杂性，并没有统一的评价标准。综合国内外文献共用的评价方式，通常会结合主观评价标准和客观评价标准来对可视化后的结果进行评价。

(1)主观评价标准

主观评价方法通过将模型注意力图展示给用户，依据他们视觉的主观评价来对算法的修复效果作出评估。由于主观评价标准受到用户心理状态、感兴趣区域等影响，不同的观测人员对同一段图像的评价可能存在差异，因此一般采用过半数以上的评价作为最终评价。

(2)客观评价标准

客观评价标准借助实验数据和评价指标来说明稳定后的图像在某方面的特性。其评价结果较为直观，可直接通过数值来比较出互学习算法可解释性的优劣。常用的评价指标包括消融实验与同类算法比较的方式，用于判定算法的有效性。

消融实验通过控制变量法，人工修改网络的某些部分或参数，以便更好地理解网络的行为，验证算法的有效性。本次实验通过展示模型分类的准确率来验证算法对模型性能的提升，利用视觉上注意力图标注的准确度来衡量模型可解释性的提升。

为了对可解释性进行量化分析，本文引入平均下降率(Average Drop，AD)与平均上升率(Average Increase，AI)作为量化指标，对注意力图的可解释性进行验证[12]。

注意力图保留了图像中特定类别显著区域的完整性，而图像的遮挡部分会降低模型在决策中的可信度。平均下降率表示遮挡后，图像中特定类的模型置信度下降的平均百分比，如式(6)所示：

$$AD=\sum_{i=1}^{N}\frac{\max(0,Y_i^c-O_i^c)}{Y_i^c}\times 100 \tag{6}$$

置信度提升可用平均上升率表示，如式(7)所示：

$$AI=\sum_{i=1}^{N}\frac{Sign(Y_i^c<O_i^c)}{N} \tag{7}$$

其中，Y_i^c 表示类别 c 在图像 i 上的预测得分，O_i^c 表示输入类别 c 在特征映射区域的预测得分，$Sign$ 表示为二分函数，如果输入为 True，则返回 1。

4.3 结果分析

(1)面向分类任务模型的有效性分析实验

本文从模型准确度方面分析了算法的有效性。其中，ST1 为预训练模型，为 ST2 提供固定的先验知识。表 1 列出了不同模型结构下全局迁移和局部迁移的准确率。

表 1 不同数据集下 ST1 模型和 ST2 模型的表现

Table 1 Performance of ST1 and ST2 models in different datasets

数据集	模型类型		全局迁移/%		局部迁移/%	
	ST1	ST2	ST1	ST2	ST1	ST2
CIFAR-10	WRN-16-2	ResNet-50	90.07	85.82	92.42	88.73
	WRN-16-2	MobileNet	90.07	82.37	92.42	85.26
CIFAR-100	WRN-16-2	ResNet-50	71.47	68.87	73.87	71.54
	WRN-16-2	MobileNet	71.47	67.52	73.87	70.54

表 1 列出了注意力图全局迁移与局部迁移对 ST1 模型和 ST2 模型分类准确率的影响。两个模型单独进行训练时，ST1 模型因具有先前教师模型知识的训练，准确率较高，而 ST2 模型未能取得良好的准确率。

在经过注意力图全局互相迁移之后，ST1 模型和 ST2 模型的准确率均有提升，有效证明了 ST2 模型学习到了 ST1 模型所学习到的知识，表明相互注意力损失函数有利于模型的性能提升，如图 5 所示。

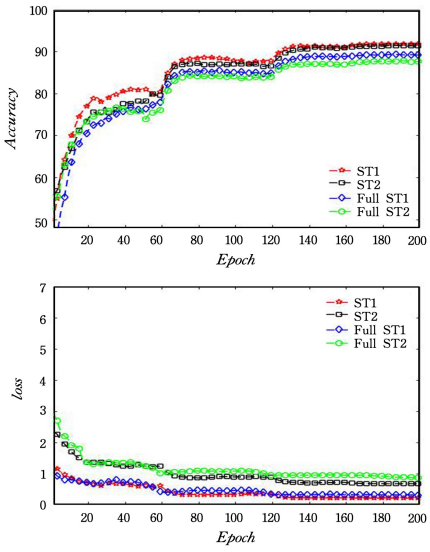

图 5 准确度与损失函数折线图

Fig. 5 Line plot of accuracy and loss function

为研究局部迁移训练策略对模型的影响，对全局迁移和局部迁移算法进行了对比。两个独立模型分别部署在不同的 GPU 上进行训练，以保证两个模型的预测值和参数值同时更新。

在 CIFAR-10 上使用 ResNet-50 和 MobileNet 进行实验，图 5 给出了损失函数收敛性和准确度的结果。两个模型在进行分布式训练时收敛更快，因为它们始终具有完全相同的训练迭代次数。由图 5 可知，所提算法有效提高了分类准确性，且 loss 函数收敛更快。

(2) 独立训练模型与注意力互学习算法的注意力图比较

本节通过 Grad-CAM 可视化经过互学习的 ST1 与 ST2 的注意图，对比未经局部互迁移模型的注意力图来验证互学习算法的有效性。在注意力图区域选取注意力图贡献度前 80%的区域由高到低进行转移。由于所使用的方法与全局迁移密切相关，因此重点对独立训练模型与全局注意力迁移模型进行比较。实验结果如图 6 所示。

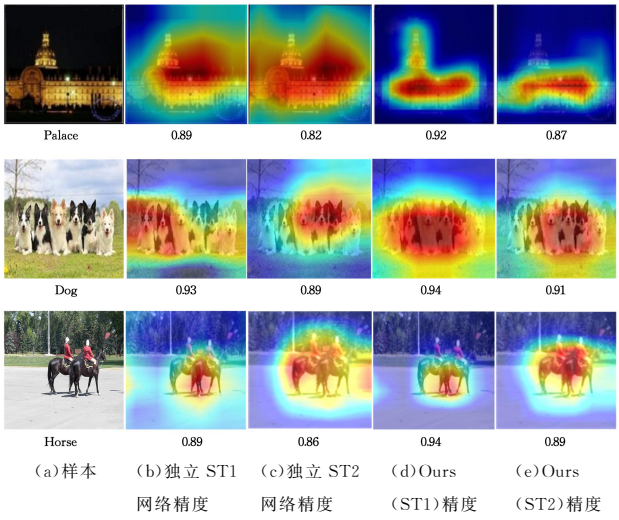

图 6 独立训练注意力图与局部迁移注意力图的对比

Fig. 6 Comparison of independent training attention maps and local transfer attention maps

图 6 为 CIFAR-10 数据集上 ST1 模型与 ST2 模型独立训练的注意图与局部互学习模型的注意力图。如图可知，独立训练模型注意力图能大致标注出模型做出决策的部分，但标注区域涵盖了错误部分，难以在视觉上作出判断。

由图 6 的实验结果可知，与独立学习相比，ST2 网络与具备先验知识的 ST1 网络进行局部注意力图迁移的方法，使模型注意力图的标注更准确，给出的决策依据明确，可解释性更高。

与此同时，为了进一步验证互迁移算法对模型内部注意力图的改进，并更好地理解模型，选取模型低层、中层与高层注意力图进行可视化，如图 7 所示。

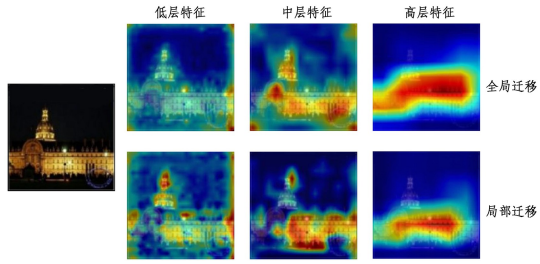

图 7 模型内部逐层注意力图对比

Fig. 7 Layer-by-layer comparison of attention maps inside model

实验结果表明，模型内部注意力图有显著优化，提升了模型内部决策过程的可解释性。

下文将对全局迁移与局部迁移给模型带来的影响进行验证，以进一步说明算法的有效性。

(3)全局迁移与局部迁移方法比较

图 8 给出了 CIFAR-10 数据集上，全局迁移注意图与局部迁移注意力图的比较结果。结果表明，在相互学习的过程中，通过局部注意力图的互学习，有效避免了全局迁移学习过程中因部分区域相似度过低而造成负迁移。

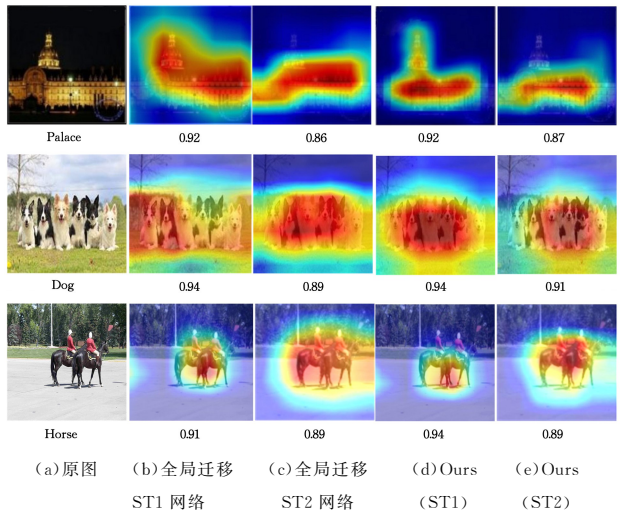

图 8 全局迁移与局部迁移注意力图的对比

Fig. 8 Comparison of attention maps between global and local transfer

例如，“Palace”的分类中，ST2 模型未经训练时，标注区域覆盖了天空部分，或是 ST1 模型未能完全覆盖需要标注的区域，导致模型注意力图标注有偏差，而局部注意力图迁移

互学习后的模型可以较好地解决这个问题。模型局部区域差异较大时，模型通过相似度度量计算出相似度较高与较低的区域，从而丢弃相似度较低的区域，避免负迁移带来的影响，反映在图上的深色区域部分，因此模型对这些区域进行学习或是互学习得到的特征权重对模型的影响较小，有效避免了负迁移对模型的影响。而标注正确的区域模型会予以保留，反映在注意力图上的高亮区域，高亮区域可反映出模型对图片做出决策的依据，并体现出模型的可解释性。

4.4 同类算法性能比较分析实验

除视觉层面可视化评估外，为了对可解释性进行量化分析，本节引入类别置信度平均下降率与平均上升率作为量化指标，对注意力图可解释性进行验证。

本节使用模型注意力图可视化可解释性置信度评估方法，主要通过将输入样本与注意力图权重作哈达玛积，使用二值化掩膜对输入样本进行遮挡，进而观察目标类上准确度的变化。在本实验中，将注意力图中 50% 的显著像素进行遮挡后与原始输入进行点乘，并使用 ImageNet 抽样所得的 ILSVRC2012 数据集进行实验，实验结果如表 2 所列（平均下降率越低，平均上升率就越高，性能就越好）。

表 2 平均下降率与平均上升率的评估结果

Table 2 Evaluation results of average drop rate and average increase rate

（单位：%）

Methods	RISE[13]	GradCAM[14]	Grad-CAM++[15]	Score-CAM[16]	Ours
AD	47.0	47.8	45.5	31.5	28.2
AI	14.0	19.6	18.9	30.6	29.5

由表 2 可知，所提算法分别实现了 28.2% 的平均下降率和 29.5% 的平均上升率，与最先进的算法 ScoreCAM 相比，其在平均下降率方面实现了 3.3% 的提升，虽然平均上升率略有下降，但总体表现良好。上述实验结果表明，所提算法能成功地找出样本图像中预测标签最相关区域，而不局限于视觉可视化区域，与现有的同类方法相比，所提方法能更准确地揭示原始 CNN 模型的决策依据。

为了探究注意力图局部互迁移对模型表现的影响，表 3 列出了近年来在深度学习视觉解释领域典型的迁移模型，以此验证所提算法的有效性。

表 3 同类算法的对比

Table 3 Comparison of similar algorithms

（单位：%）

	CIFAR-10	CIFAR-100
KD($T=1$)[9]	86.37	69.13
DML[17]	88.10	**72.73**
UL-H[18]	84.25	68.53
KDFM[19]	87.51	70.10
Ours	**88.73**	71.54

本文算法与最先进的算法 DML 相比，在 CIFAR-10 数据集上准确率实现了 63% 的提升，虽然在 CIFAR-100 数据集上分类准确率略有下降，但有效提升了注意力图标注的准确性，在可解释性上优于最先进方法。

上述实验结果表明，互迁移算法在模型可解释性方面有显著提升。该算法对模型准确率的提升，在一定程度上表明特征图显示了模型提取出的特征本质规律。具体表现为：模型的泛化能力越强，模型就越有可能挖掘到权重的内在特性，并通过特征图表现。模型除了学习到正确的特征分类，还能学习到错误分类与正确分类共有的一些特征，使得模型在测试数据上更有可能捕捉更多特征的多种特性，表现出较强的泛化能力，有效改进了特征图权重分布，反映在注意力图上则标注区域更为精准。因此，模型之间进行注意力图交互有利于学习到各自的权重分布特性，从而改善模型的泛化性能。

结束语 本文针对独立模型构建出的注意力图存在的标注区域不准确而导致可视化可解释性较差的问题，提出了一种基于注意力图的局部互迁移方法。通过叠加特征图得到注意力图后，对注意力图进行区域划分，以度量不同区域相似度，并对贡献度较高且相似度较大的区域进行迁移，以提高注意力图标注区域的准确度，避免负迁移带来的不良效果，从而提升模型的可视化可解释性。本文在可解释性深度学习研究方面取得了初步进展，未来将致力于突破基于结果论进行解释的局限性。

参 考 文 献

[1] YUAN X X, WU Q. Object detection in remote sensing images based on saliency feature and angle information[J]. Computer Science, 2021, 48(4): 174-179.

[2] YADAV A, VISHWAKARMA D K. Sentiment analysis using deep learning architectures: a review[J]. Artificial Intelligence Review, 2020, 53(6): 4335-4385.

[3] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// Proceedings of European Conference on Computer Vision. Berlin: Springer, 2014: 818-833.

[4] GAUR M, FALDUK, SHETH A. Semantics of the black-box: can knowledge graphs help make deep learning systems more interpretable and explainable? [J]. IEEE Internet Computing, 2021, 25(1): 51-59.

[5] DENIL M, BAZZANI L, LAROCHELLE H, et al. Learning where to attend with deep architectures for image tracking [J]. Neural Computation, 2012, 24(8): 2151-2184.

[6] 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.

[7] 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.

[8] ZENG T, ACUNA D E. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models[J]. Scientometrics, 2020, 5(1): 1-30.

[9] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.

[10] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: fast optimization, network minimization and transfer learning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4133-4141.

[11] HSU E Y, LIU C L, TSENG V S. Multivariate time series early classification with interpretability using deep learning and attention mechanism[M]. Advances in Knowledge Discovery and Data Mining. 2019.

[12] LIU Q, MUKHOPADHYAY S. Unsupervised learning using pretrained CNN and associative memory bank[C]// 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-8.

[13] PETSIUK V, DAS A, SAENKO K. Rise: randomized input sampling for explanation of black-box models [EB/OL]. (2018-09-25) [2018-09-25]. https://arxiv.org/abs/1806.07421.

[14] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[15] CHATTOPADHYAY 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 (WACV). IEEE, 2018: 839-847.

[16] 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.

[17] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4320-4328.

[18] LIU Q, MUKHOPADHYAY S. Unsupervised learning using pretrained CNN and associative memory bank[C]// 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-8.

[19] CHEN W C, CHANG C C, LEE C R. Knowledge distillation with feature maps for image classification[C]// Asian Conference on Computer Vision. Cham: Springer, 2018: 200-215.

CHENG Ke-yang, born in 1982, Ph. D, professor, is a member of China Computer Federation. His main research interests include computer vision and machine learning.

(责任编辑: 李亚辉)