雷达图像深度学习模型的可解释性研究与探索

李伟杰, 杨威*, 刘永祥, 黎湘

国防科技大学电子科学学院, 长沙 410073

* 通信作者. E-mail: yw850716@sina.com

收稿日期: 2021-03-22; 修回日期: 2021-05-09; 接受日期: 2021-06-07; 网络出版日期: 2022-06-10

国家自然科学基金 (批准号: 61871384, 61401486)、国家创新研究群体科学基金 (批准号: 61921001) 和湖南省自然科学基金 (批准号: 2017JJ3367) 资助项目

摘要 虽然深度学习技术在雷达图像处理任务中获得了部分应用, 但缺乏对黑盒模型的可解释性分析和全面的性能评估, 限制了该技术在雷达图像领域中的应用性能、可信度和广泛性. 本文从可解释性入手, 提出了雷达图像深度学习黑盒模型分析思路, 在开源 MSTAR 雷达图像数据集上进行实验验证. 从深度学习模型的迁移机制和认知机理两个方面进行分析, 得到了关于迁移学习、归因方法应用和模型鲁棒性评估方面的相关结论, 填补了现有研究的空白.

关键词 深度学习, 雷达图像, 可解释性, 迁移学习, 归因方法, 鲁棒性

1 引言

近些年来, 深度学习技术正逐步从光学图像领域应用到雷达图像领域, 在雷达目标识别、场景分类和目标检测等雷达图像处理任务中取得了部分研究成果^[1,2]. 目前深度学习模型在以上任务中表现出了优良性能, 但仍存在一些尚待研究的问题.

一是需要根据雷达与光学图像的显著差异性, 进一步解释深度学习模型在雷达图像领域中的迁移机制, 指导雷达图像深度学习模型的优化设计. 由于雷达与光学成像机理不同, 所得到的图像为目标散射点分布. 如图 1 所示, 在 MSTAR 数据集^[3]中表现为呈离散分布、强度不同的散射点, 与所对应的光学图像之间存在明显差异. 将在光学图像领域中取得成功应用的深度学习模型迁移到雷达图像领域时, 应当充分考虑这种差异性. 文献 [2] 梳理了小样本条件下基于迁移学习的雷达图像目标识别方法, 部分研究针对两者的差异性设计了相应方法提高目标识别率, 但缺乏对其迁移机制影响的深入分析, 可解释性研究有望为深度学习模型在雷达图像领域中的设计与优化提供全新的视角和知识.

二是由于雷达图像目标之间、目标与背景之间具有较强相似性, 需要对深度学习模型认知机理的分析评估, 保证其在实际任务中的应用性能和可信度. 如图 1 所示, 这种相似性给人的认知和理解带

引用格式: 李伟杰, 杨威, 刘永祥, 等. 雷达图像深度学习模型的可解释性研究与探索. 中国科学: 信息科学, 2022, 52: 1114–1134, doi: 10.1360/SSI-2021-0102
Li W J, Wang W, Liu Y X, et al. Research and exploration on the interpretability of deep learning model in radar image (in Chinese). Sci Sin Inform, 2022, 52: 1114–1134, doi: 10.1360/SSI-2021-0102

图 1 (网络版彩图) MSTAR 数据集的雷达图像与对应的光学目标. (a) 2S1; (b) BMP2; (c) BRDM2; (d) BTR60; (e) BTR70; (f) T72

Figure 1 (Color online) Radar image and corresponding optical target in MSTAR dataset. (a) 2S1; (b) BMP2; (c) BRDM2; (d) BTR60; (e) BTR70; (f) T72

来了极大的困难, 需要分析模型是否正确表征了目标信息. 目前对于模型的分析评估大多依赖于识别率, 对实际任务而言, 识别率是对问题的不充分描述, 这可能会使得模型出现错误, 例如模型利用背景信息的相关性提高识别率. 文献 [4] 提到华盛顿大学 (University of Washington) 的一项实验, 旨在创建哈士奇与狼的识别器, 但模型实际上是根据背景是否存在积雪来进行识别的, 学习到的是背景干扰与识别结果的强相关性, 而不是目标特征与识别结果之间的因果性. 文献 [5] 同样指出, 在测试集上的高识别率并不能保证模型学习到正确的特征表征. 仅靠识别率等性能指标不能保证模型在实际任务中正常工作, 也不能满足在某些关键任务 (如医疗、交通和军事等) 中人们的认知和决策需要. 因此, 只有对深度学习模型认知机理的分析说明和识别结果的可信解释, 才能提高人们对于其黑盒模型的信任, 保证人们依靠其作出相应决策.

针对以上两个问题, 本文将使用可解释性方法^[6,7]作为分析工具从两方面展开研究. 一是通过可解释性方法对比分析雷达和光学图像对深度学习模型的影响, 获取其黑盒模型内部知识, 解释深度学习模型迁移机制, 从而进一步理解应用深度学习技术. 二是通过主观分析和客观指标评估可解释性方法和深度学习模型, 从而在准确性和鲁棒性方面得到可信结论, 以此对其黑盒模型的认知机理作出分析评估, 为实际决策提供帮助.

目前可解释性方法可分为事后解释 (post-hoc) 方法和事前设计 (ante-hoc) 方法. 事后解释方法针对已经训练好、不具备可解释性的黑盒模型进行研究, 通过可视化展示模型内部结构、基于归因方法 (attribution algorithms) 分析模型识别重点、使用特定输入研究模型、比较不同模型结构性能等方式, 挖掘模型内部知识和解释模型认知机理. Zeiler 等^[8]提出反卷积 (deconvnet) 的分析方法. Simonyan 等^[9]提出两种基于梯度的分析方法. Szegedy 等^[10]通过单元分析方法探索神经元的语义信息, 并发现了对抗样本. Frankle 等^[11]提出乐透假设 (lottery ticket hypothesis) 发现训练成功的大型网络中包含一个相同性能的稀疏子网络.

事前设计方法构建具有明确语义或物理信息的模型, 可将数学物理上具有明确意义的方法与深度学习模型相结合, 提高模型的可解释性. 如通过知识蒸馏结合神经网络和结构化的逻辑规则^[12]; 利用注意力机制同时生成图像和文本解释^[13]; 贝叶斯 (Bayes) 深度学习将深度学习模型的感知能力和贝叶斯模型的因果逻辑推理能力相结合^[14]; Gu 等^[15]基于最优传输理论得到了半透明的生成对抗网

络. 事前设计方法也可设计本身在语义、物理等层面上具有可解释性的模型. Sabour 等^[16] 提出胶囊网络 (capsule network) 模型, 胶囊单元不同维度表征了手写数字的语义信息; Zhang 等^[17] 设计了一种带有可解释性的卷积核.

当前, 有关雷达图像领域的深度学习可解释性方面研究较少. 郭炜炜^[18] 讨论了关于合成孔径雷达 (synthetic aperture radar, SAR) 图像目标识别的可解释问题, 从模型理解、模型诊断和模型改进等方面进行了初步探讨, 但暂无对雷达图像深度学习黑盒模型可解释性的具体实验研究.

本文结合目前深度学习可解释性方面的研究进展, 以雷达目标识别任务为例, 在 MSTAR 的 SAR 图像数据集上进行事后解释方法的实验分析, 实验结果有效揭示了深度学习模型在雷达图像领域中的迁移机制和认知机理, 并且详细地提供了相应的分析思路和结论, 填补了当前对雷达图像深度学习黑盒模型可解释性的研究空白. 主要贡献如下:

(1) 基于事后分析方法提出了雷达图像深度学习黑盒模型分析思路, 为可解释性方法在雷达图像领域中的应用奠定基础;

(2) 通过实验比较分析了雷达和光学图像对于模型内部的影响, 得到了关于深度学习在雷达图像领域中迁移机制的分析结论;

(3) 通过归因方法分析了深度学习黑盒模型的认知机理, 从可解释性角度提出了 SAR 目标识别模型鲁棒性评估的新方法.

本文后续结构组织如下: 第 2 节介绍相关工作, 第 3 节面向雷达图像领域提出深度学习黑盒模型分析思路, 第 4 节依据所提思路进行实验验证, 第 5 节对全文总结并提出下一步研究方向.

2 相关工作

由于本文基于事后解释方法分析黑盒模型, 而其中的归因方法为当前主要研究方向, 能有效地分析模型输入与输出之间的联系, 下面主要介绍该方法.

归因方法将输入与输出联系起来, 为输入提供一个与之相对应的权重. 设输入空间为 $X \subseteq \mathbb{R}^m$, 输出空间为 $Y \subseteq \mathbb{R}^n$, 函数空间为 $F: \mathbb{R}^m \mapsto \mathbb{R}^n$, 归因方法可定义为 $\phi: F \times \mathbb{R}^m \mapsto \mathbb{R}^m$, 对于给定的输入 x , 输出 $f(x)$, 其中 $f \in F$ 表示黑盒模型. 归因方法会给出与输入相对应的重要性分数 $\phi(f, x)$ 作为显著性图 (saliency/importance/attribution map). 可根据输入输出的对象不同分为 3 类: 主要归因 (primary attribution) 方法、层归因 (layer attribution) 方法和神经元归因 (neuron attribution) 方法^[19].

2.1 主要归因方法

主要归因方法关注黑盒模型整体, 计算模型输入对于模型输出的贡献作为显著性图, 衡量的是输入特征对于输出结果的重要性. 由于梯度衡量输出对于输入的灵敏程度, 可使用输出 $f(x)$ 对输入向量 x 的梯度 $[\nabla_x f(x)]_i$ 作为输入第 i 个分量的贡献衡量指标. Simonyan 等^[9] 提出通过一次梯度反向传播得到输入图像像素点对应的梯度作为显著性图, 其方法可视为一阶泰勒近似, 梯度作为线性模型中的系数, 输入乘以梯度相当于特征对于线性模型输出的贡献. 该方法为文献 [8] 中反卷积方法的推广. Springenberg 等^[20] 在文献 [8, 9] 的基础上提出导向反向传播 (guided-backpropagation, GB), 用于分析对网络结果有正面贡献的部分. 文献 [21] 提出 Guided Grad-CAM (guided gradient-weighted class activation mapping, GGC) 结合了 GB 的高分辨率和 Grad-CAM 的类映射优势.

值得指出的是, 以上梯度反传方法存在梯度饱和和梯度消失等现象^[22], 因此一系列改进算法被

相继提出. Bach 等 [23] 提出 LRP (layer-wise relevance propagation) 逐层计算相关性分数 (relevance score), 最终得到输入与输出的相关性分数 $R_p^{(1)}$ (pixel-wise relevance score). 该方法重新定义了反向传播规则, 对每层模型作线性近似, 每一层的相关性分数是上一层的加权求和, 有 ε -rule 和 β -rule 两种计算规则. Binder 等 [24] 后续利用一阶泰勒 (Taylor) 展开将 LRP 拓展到非线性层. Shrikumar 等 [22] 提出 DeepLIFT (deep learning important features), 引入参考点概念, 对输入 x 分配的贡献为 $C_{\Delta x \Delta t}$, 参考输入差异 Δx 和参考输出差异 Δt 之间的关系满足 $\sum_{i=1}^n C_{\Delta x_i \Delta t} = \Delta t$, 而 $m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}$ 作为梯度, 定义相应的反向传播规则进行计算. 重新考虑文献 [9] 中的梯度反传方法, 对于模型在参考点 x' 处作一阶泰勒展开, 则有 $f(x) - f(x') \approx \sum_i [\nabla_x f(x)]_i [x - x']$, 基于梯度的方法可视作 x' 和 $f(x') = 0$ 的特殊情况. 引入参考点后的重要性是对参考样本相对重要程度的衡量. 基于此, Sundararajan 等 [25] 提出积分梯度 (integrated gradients, IG), 通过参数曲线 $\gamma(\alpha)$ 连接 x 和 x' , 其中 $\gamma(0) = x$, $\gamma(1) = x'$, 由定积分运算得

$$f(x) - f(x') = \int_0^1 \langle \nabla_x f(\gamma(\alpha)), \gamma'(\alpha) \rangle d\alpha = \sum_i \int_0^1 [\nabla_x f(\gamma(\alpha))]_i [\gamma'(\alpha)]_i d\alpha. \quad (1)$$

在参数曲线为直线 $\gamma(\alpha) = (1 - \alpha)x + \alpha x'$ 的情况下对输入第 i 个特征有

$$\text{IntegratedGrads}_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial f(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha. \quad (2)$$

文献 [25] 进一步讨论了归因方法应当满足敏感性 (sensitivity) 和不变性 (implementation), 如 IG. 基于梯度反传的方法破坏了敏感性, LRP 和 DeepLIFT 不满足不变性. Ancona 等 [26] 基于文献 [22, 25] 定义了广义敏感性 (sensitivity-n), 并讨论了上述 LRP, DeepLIFT 和 IG 等基于梯度的方法之间形式上的联系.

另一个研究方向是对输入进行扰动, 观察输出的变化. 文献 [8] 利用灰色矩形对彩色图像进行遮挡, 也可使用参考值 (如 0 等) 替换样本特征值. 特征置换 [27] (feature permutation) 通过随机排列一批样本特征值顺序进行扰动, 文献 [28] 建议将数据集分成两半并交换对应特征值. LIME [29, 30] (local interpretable model-agnostic explanation) 通过对输入样本的局部扰动构造新数据集, 并在此基础上训练可解释性模型 (如线性模型) 近似原模型. 但基于扰动的方法计算效率低于基于梯度的方法, 并且 LIME 的超像素分割结果在 SAR 图像中会包含目标和相似背景. 扰动也可用于提高显著性图质量, SmoothGrad [31] 和 VarGrad [32] 通过对输入添加高斯 (Gauss) 噪声, 实现多次采样以锐化显著性图, 如式 (3) 所示计算采样均值, 如式 (4) 所示计算采样方差.

$$E(\phi(x, \xi_k)) = \frac{1}{n} \sum_{k=1}^n \phi(x, \xi_k), \quad \xi \sim \mathcal{N}(0, \sigma^2), \quad (3)$$

$$\text{Var}(\phi(x, \xi_k)) = \frac{1}{n} \sum_{k=1}^n \{\phi(x, \xi_k)\}^2 - \{E(\phi(x, \xi_k))\}^2, \quad \xi \sim \mathcal{N}(0, \sigma^2). \quad (4)$$

从博弈论的沙普利值 [33] (shapley value) 出发, 也有相关研究以此衡量贡献值. 根据联盟博弈理论 [34], 沙普利值是满足效率、边际性和对称性的唯一单值解, 因此具有衡量贡献的良好性质. 文献 [35, 36] 利用蒙特卡洛 (Monte Carlo) 采样估计沙普利值. 文献 [37] 提出线性可加性特征归因方法 (additive feature attribution methods) 和 SHAP (shapley additive explanations) 值, SHAP 值是满足局部准确性 (local accuracy)、无效性 (missingness) 和一致性 (consistency) 的唯一解, 其中无效性是对于沙普利值的改进. 对于 SHAP 值计算有多种方法, 文献 [37] 结合 LIME 提出了 KernelSHAP. 针对深度学习模型, 文献 [37] 结合 IG 和 DeepLIFT 分别提出了 Gradient SHAP 和 DeepSHAP.

2.2 层归因方法

层归因方法关注黑盒模型内部对于输出的影响, 计算模型内部特定层输出对模型输出的贡献作为显著性图, 可用于分析在识别过程中层的重要性. 通过将层输出代替模型输入, 上述基于梯度和扰动的方法可应用于此. Dhamdhere 等^[38]在IG的基础上提出了层电导系数(layer conductance), 某层神经元输出 y 的层电导系数为 $\text{Cond}^y(x) = \sum_i (x_i - x'_i) \cdot \int_{\alpha=0}^1 \frac{\partial f(x' + \alpha \times (x - x'))}{\partial y} \cdot \frac{\partial y}{\partial x_i} d\alpha$. 文献[39]同样使用类似IG积分的思想, 根据层对梯度进行积分衡量层内部影响(internal influence). Grad-CAM^[21]和 Grad-CAM++^[40]等是用于卷积神经网络(convolutional neural networks, CNN)的层归因方法, 通常在最后一层卷积层的特征图上分析, 由于卷积层输出与输入图像在空间位置上存在对应关系, 因此可上采样到原图大小.

2.3 神经元归因方法

神经元归因方法关注输入特征和黑盒模型内部特定神经元之间的关系, 计算输入特征对某层神经元输出的贡献作为显著性图. 通过将神经元输出代替模型输出, 上述基于梯度和扰动的方法也可应用于此. 文献[8]的反卷积网络可用于分析卷积层的语义信息. 在文献[38]中, 某层神经元输出 y 对输入特征 x_i 的电导系数定义为 $\text{Cond}_i^y(x) = (x_i - x'_i) \cdot \int_{\alpha=0}^1 \frac{\partial f(x' + \alpha \times (x - x'))}{\partial y} \cdot \frac{\partial y}{\partial x_i} d\alpha$, 由此可分析神经元输出对于输入特征的识别重点.

3 雷达图像深度学习黑盒模型分析思路

现有研究并未系统地对雷达图像领域中深度学习黑盒模型进行可解释性分析, 本节由此尝试提出雷达图像深度学习黑盒模型分析思路. 3.1 小节介绍基于事后解释方法对迁移机制的分析思路. 3.2 小节介绍基于主要归因方法对模型认知机理的分析思路, 并引入了客观指标对其进行评估.

3.1 基于事后解释方法的模型迁移机制的对比分析

根据前文所述, 在黑盒模型内部, 从以下3条途径分析雷达和光学图像差异性对迁移机制的影响.

(1) 基于可视化的卷积核和特征图分析. 通过可视化卷积核权重和特征图, 直观地分析在不同数据类型下模型卷积核的特点和对目标特征的提取能力.

(2) 基于非线性降维方法的可视化分析. t-SNE^[41], UMAP^[42]等非线性降维方法可用于分析数据在模型各层上的分布特点. 但由于非线性降维方法不可避免地造成了信息损失, 同时神经网络本身就是一种良好的降维方法, 其降维的结果直接用于分类识别. 故修改全连接层输出至二维作可视化分析.

(3) 基于层归因方法和神经元归因方法的模型内部分析. 层归因方法分析识别过程中哪些神经元对输出起到关键作用, 神经元归因方法分析不同神经元输出对于输入特征的识别重点, 并使用层电导系数分析底层卷积核的特性.

3.2 基于主要归因方法的模型认知机理分析评估

主要归因方法将模型视作一个整体, 从单个样本的局部角度解释模型认知机理. 使用了多种主要归因方法对多种模型进行分析评估, 以得到可信的分析结论.

在模型理解方面, 通过显著性图分析模型的识别重点, 解释数据相似性对其认知机理的影响, 以此判断模型是否学习到了输入和输出之间的因果性, 是否符合人类认知逻辑和物理规律.

在模型诊断方面,通过显著性图的异常分析模型的潜在缺陷.基于显著性图进一步诊断模型对数据特征的学习情况,从不同角度研究模型的泛化性能.

在模型评估方面,通过不同模型显著性图之间的差异,研究模型对重要特征的提取和认知能力,从而比较评估模型性能.

使用主要归因方法进行主观分析的同时,需要客观指标评估主要归因方法.在光学图像领域中,可基于人工标注信息评估显著性图,但人工标注信息较难获取.此外,与光学图像不同, SAR 目标识别模型利用了背景相关信息辅助识别,根据标注信息评估主要归因方法时会受到此因素的较大影响.故本文采用文献 [43] 的失真度 (infidelity) 和最大灵敏度 (max-sensitivity) 对主要归因方法进行评估.基于文献 [25, 26] 的敏感性指标,文献 [43] 定义了失真度反映主要归因方法的准确性,对输入施加扰动,主要归因方法应准确反应模型输出的相应变化.式 (5) 计算服从分布 μ_I 的输入扰动 I 乘显著性图 $\phi(f, x)$ 和输出变化的均方误差,该值越小说明主要归因方法准确性越好.

$$\text{Infidelity}(\phi, f, x) = \mathbb{E}_{I \sim \mu_I} [(I^T \phi(f, x) - (f(x) - f(x - I)))^2]. \quad (5)$$

对于主要归因方法的鲁棒性评估使用最大灵敏度 [43].在领域半径为 r 的扰动下,式 (6) 通过蒙特卡洛采样计算显著性图的最大变化,该值越小说明主要归因方法鲁棒性越好.

$$\text{Max-Sensitivity}(\phi, f, x, r) = \max_{\|y-x\| \leq r} \|\phi(f, y) - \phi(f, x)\|. \quad (6)$$

文献 [44] 发现模型的鲁棒性越强,其显著性图与输入图片的一致性越好,即更符合人的主观认识.对于 SAR 目标识别模型而言,可认为模型的鲁棒性和其显著性图中对于背景干扰信息的去除能力有关.因此,基于失真度和最大灵敏度指标选择主要归因方法后,利用显著性图中背景干扰成分来近似衡量模型鲁棒性.当有人工标注信息时,可以此计算标注信息和二值化显著性图的交并比 (intersection over union, IoU), IoU 计算公式为

$$\text{IoU} = \frac{\text{TP}}{\text{FP} + \text{TP} + \text{FN}}, \quad (7)$$

其中 TP, FP 和 FN 分别表示真正例 (true positive), 假正例 (false positive) 和假负例 (false negative), 该值越大说明深度学习黑盒模型鲁棒性越好.

但人工标注成本较高,在缺乏标注信息的情况下.如图 2 所示,我们提出了一种基于 SmoothGrad 和 VarGrad 的模型鲁棒性评估方法.根据文献 [45] 的研究,当输入扰动为高斯噪声时,SmoothGrad 的结果包含显著性图、高阶偏导数级数和高斯噪声的标准差,VarGrad 的结果包含高阶偏导数级数和高斯噪声的标准差,两者之间差异主要在于显著性图部分.因此,使用 SmoothGrad 和 VarGrad 计算 IoU 近似衡量模型鲁棒性.图 3 的实验结果说明了该方法的合理性.如图 3 所示,每张输入图像第 1 行为 SmoothGrad,第 2 行为 VarGrad,从左至右为第 4 节中不同模型,VarGrad 的显著性图包含背景干扰远少于 SmoothGrad.可以认为,目标信息包含于显著性图 (即一阶偏导数) 和高阶偏导数级数中,而背景干扰主要包含于显著性图中.因此,将 VarGrad 计算得到的显著性图作为基准标注信息与 SmoothGrad 进行计算近似衡量显著性图中包含的背景信息,所得到的 IoU 值越大说明包含的背景信息越少,即模型鲁棒性越好.

根据算法 1 将 SmoothGrad 和 VarGrad 的显著性图结果经过二值化后计算 IoU.不同于在测试集上验证泛化性能的识别率评估方法,该方法从可解释性角度评估了模型对于目标因果信息的提取能力,克服了测试集数据不足和与训练集差异较小的问题,并且符合人的逻辑认识,有助于为决策提供依据.

图 2 (网络版彩图) 基于 SmoothGrad 和 VarGrad 的 IoU 计算示意图

Figure 2 (Color online) Schematic diagram of IoU calculation based on SmoothGrad and VarGrad

图 3 (网络版彩图) 基于 SmoothGrad 和 VarGrad 的 IG 显著性图

Figure 3 (Color online) IG saliency map based on SmoothGrad and VarGrad

4 实验与分析

4.1 实验说明

1120 本文在 MSTAR 的 SAR 数据集上进行实验, SAR 数据集为 128×128 的坦克图像, 数据集构成如

算法 1 IoU computing based on SmoothGrad and VarGrad**Input:** Data $\{x_1, x_2, \dots, x_n\}$, model f , attribution algorithm ϕ , sample t , stdev σ .**Output:** IoU.

```

1: for  $i = 1$  to  $n$  do
2:   for  $k = 1$  to  $t$  do
3:     Obtain  $\phi(x_i, f, \xi_k)$  based on Gaussian noise sampling  $\xi_k \sim \mathcal{N}(0, \sigma^2)$ ;
4:   end for
5:   Compute  $\phi(x_i, f, \xi_k)$  mean value SmoothGrad and  $\phi(x_i, f, \xi_k)$  variance VarGrad;
6:   SmoothGrad and VarGrad results are converted to the binary;
7:   Calculate IoU from binary results;
8: end for
9: Take the mean of IoU on the Data.

```

表 1 MSTAR 数据集**Table 1** MSTAR dataset

Target type	Training set (17° pitch angle)	Test set (15° pitch angle)
2S1	299	274
BMP2	232	196
BRDM2	298	274
BTR70	256	195
BTR60	233	196
D7	299	274
T62	299	273
T72	232	196
ZIL131	299	274
ZSU234	299	274

表 2 模型训练结果**Table 2** Model training results

Model	Training set (%)	Test set (%)
Three-layer CNN (random initialization)	100.0	97.24
VGG-16 (random initialization)	99.96	95.47
VGG-16 (pre-trained)	100.0	98.43
ResNet-50 (random initialization)	100.0	96.37
ResNet-50 (pre-trained)	100.0	98.10

表 1 所示. 对于 SAR 单通道灰度图像, 将其从单通道复制为三通道以适配光学图像模型.

使用的模型有三层卷积块 (Conv2d-MaxPool2d-BatchNorm2d-ReLU) 加二层全连接层的卷积神经网络、VGG-16^[46] 和 ResNet-50^[47]. 在雷达图像上直接训练的模型为随机初始化, VGG-16 和 ResNet-50 的迁移学习模型使用了光学数据集 ImageNet 预训练权重作为初始化, 微调过程并未冻结各层权重. 并根据输入图像大小调整了上述模型的全连接层参数. 模型识别率如表 2 所示.

图 4 第 1 层部分卷积核权重. (a) 三层 CNN (随机初始化); (b) ResNet-50 (随机初始化); (c) ResNet-50 (预训练); (d) ResNet-50 (ImageNet)

Figure 4 The weights of convolution kernel in the first layer. (a) Three-layer CNN (random initialization); (b) ResNet-50 (random initialization); (c) ResNet-50 (pre-trained); (d) ResNet-50 (ImageNet)

4.2 模型迁移机制的对比分析

本小节主要从深度学习黑盒模型内部分析雷达和光学图像差异性对迁移机制的影响, 4.2.1 小节主要分析底层卷积核, 4.2.2 小节主要分析各层特征图, 4.2.3 小节分析不同源数据的影响.

4.2.1 模型的底层机制分析

如图 4 所示, 可视化三层 CNN 和 ResNet-50 第 1 层尺寸为 7×7 的卷积核权重 (VGG-16 第 1 层尺寸为 5×5 , 因此不作可视化). 图 4(a) 和 (b) 反映使用雷达图像直接训练的底层卷积核无法学习到明确的纹理特征, 是杂乱无章的散射点分布, 这与雷达图像的散射点分布相符合. 图 4(d) 说明光学图像比雷达图像更富含纹理信息, 对应卷积核纹理特征更为明显. 文献 [20] 卷积核可视化结果和文献 [8] 反卷积结果同样印证了光学图像富含纹理信息这一观点. 图 4(c) 和 (d) 表明在训练过程中, 底层卷积核参数变化较小, 预训练的 ResNet-50 在雷达图像上微调后仍在底层卷积核保留原来的纹理特征.

图 5 第 1 层部分卷积特征图. (a) 三层 CNN (随机初始化); (b) ResNet-50 (随机初始化); (c) ResNet-50 (预训练); (d) ResNet-50 (ImageNet)

Figure 5 Partial convolution feature maps in the first layer. (a) Three-layer CNN (random initialization); (b) ResNet-50 (random initialization); (c) ResNet-50 (pre-trained); (d) ResNet-50 (ImageNet)

如图 5 所示, 输入一张 BTR70 SAR 图像, 分析第 1 层卷积核提取的特征图. 图 5(a) 和 (b) 说明直接训练的模型提取了目标整体轮廓, 并且更深的网络提取效果更明显. 图 5(c) 和 (d) 说明, 迁移后只有部分卷积核提取了整体轮廓, 其余提取的是杂乱纹理. 这说明对雷达图像而言, 仅存在部分与光学对应的有效纹理. 为进一步说明该现象, 使用了层电导系数对第 1 层卷积层进行分析, 将得到的结果取绝对值后按每个特征图求和, 各个卷积核特征图对模型输出的重要性分布如图 6 所示, 预训练的模型分布结果更接近零, 反应了只有少数卷积核输出对结果起到关键作用, 这表明数据的差异性导致预训练模型存在更多冗余.

4.2.2 模型的高层特性分析

可视化各模型从低到高的平均特征图, 如图 7 所示, 同图 5 输入同一张 BTR70 SAR 图像, 各模型在高层都捕捉到了目标的位置信息. 尽管预训练的 ResNet-50 识别率更高, 但是边缘部分噪声更多, 这可能是模型对背景干扰的过拟合. 图 7(d) 的 ImageNet 的光学模型提取到了目标信息, 说明光学纹

图 6 层电导系数分布. (a) 三层 CNN (随机初始化); (b) ResNet-50 (随机初始化); (c) ResNet-50 (预训练)
Figure 6 Layer conductance distribution. (a) Three-layer CNN (random initialization); (b) ResNet-50 (random initialization); (c) ResNet-50 (pre-trained)

理提取机制适用于不同源的数据. Huang 等^[48] 讨论了迁移学习中基于网络特征层的一般性和特殊性对不同层的选择和处理, 在网络中底层特征表现出更强的一般性, 而高层特征则具有更强的特殊性, 其实验结果显示底层特征的一般性有助于提高迁移学习模型的识别率. 结合本文分析, 底层特征的一般性和特征提取有关, 经过大量光学图像训练的底层纹理提取机制能有效提取信息, 并且符合人的逻辑认识, 因此具有明确语义信息的卷积核可能具有更强的一般性. 而高层特征与分类有关, 受到数据特殊性影响较大, 因此需要目标域数据训练.

此外, 文献 [48] 使用 t-SNE 降维可视化模型各层, 由于 t-SNE 降维过程中会损失部分信息, 因此本文参考文献 [49], 利用全连接层二维可视化, 结果如图 8 所示. 与文献 [49] 在 MNIST 数据集上结果类似, 使用 Softmax 交叉熵损失训练, 结果为条带状呈不同角度的近似均匀分布. 同时图 8 说明了迁移学习的作用, 预训练模型的特征更为集中, 类内距离更小, 在训练集上接近于文献 [49] 所提的中心损失 (centerloss) 中成团分布的效果.

Marmanis 等^[50] 在 VHR (very high resolution) SAR 数据集上同样训练随机初始化和预训练这两种方式的 ResNet-50, 其实验结果显示识别率没有明显差异. 文献 [50] 认为, 预训练权重对于模型训练没有明显帮助; 文献 [2] 认为由于光学和 SAR 数据具有较大的差异, 很难将由光学数据训练的网络直接用于 SAR 目标分类任务, 即使低层网络特征也难以进行有效的迁移. 而本文的实验结果说明, 深度学习的迁移机制在于其光学底层纹理提取机制具有较强一般性, 部分卷积核可以提取雷达图像的有用信息, 并且增强了样本可分性. 但数据的差异性导致迁移学习存在冗余性. 同时实验结果说明雷达图

图 7 输入为 SAR 图像的部分卷积层特征图. (a) 三层 CNN (随机初始化); (b) ResNet-50 (随机初始化); (c) ResNet-50 (预训练); (d) ResNet-50 (ImageNet). (a) 中最后一张为输入 BTR70

Figure 7 Partial convolution feature maps. (a) Three-layer CNN (random initialization); (b) Resnet-50 (random initialization); (c) Resnet-50 (pre-trained); (d) Resnet-50 (ImageNet). The last image in (a) is the input BTR70

像的特征提取机制并不唯一, 直接训练的呈离散点状分布特征和光学预训练的明确语义纹理特征同样有效.

4.2.3 多样化数据源对模型特征提取能力的影响

如图 9 所示, 在光学图像上进一步分析模型各层特征图. 尽管图 9(b) 模型是在雷达图像上训练的, 但依然提取到了有效的信息, 这说明对于光学和雷达图像而言, 特征提取方法并不唯一, 训练的底层卷积核具有较强的一般性. 同时预训练的 ResNet-50 提取区分度更好, 图 9(c) 从猫和花中提取花, 而其他模型均有一定的混叠. 文献 [48] 得到了相似结论, 即采用更多的数据源能取得更好的识别效果. 这说明多样化的数据 (包含光学、雷达等来源) 可能暗含不同的特征提取机制, 有助于提高模型的特征提取能力. 受限于硬件资源, 在小型光学数据集上对上述推测进行验证, 使用图 9(c) 和 (d) 的两种模型作为预训练初始化权重在光学数据集 CIFAR-10 上训练并测试, 图 9(c) 相较于图 9(d) 识别率提高了 3.7%.

图 8 (网络版彩图) 二维全连接层特征可视化. (a) ResNet-50 训练集结果 (随机初始化); (b) ResNet-50 训练集结果 (预训练); (c) ResNet-50 测试集结果 (随机初始化); (d) ResNet-50 测试集结果 (预训练)

Figure 8 (Color online) Feature visualization of 2D fully connected layer. (a) ResNet-50 in training set (random initialization); (b) ResNet-50 in training set (pre-trained); (c) ResNet-50 in test set (random initialization); (d) ResNet-50 in test set (pre-trained)

4.3 模型认知机理的分析评估

本小节基于主要归因方法从模型整体对其认知机理进行分析评估, 4.3.1 小节对方法和模型进行了定性分析, 4.3.2 小节根据定量指标进行客观评估.

4.3.1 基于主要归因方法的认知机理分析

由于基于扰动的方法计算效率不高, 本小节采用表 3 所示的主要归因方法绘制显著性图. 对三通显著性图平均得到单通道结果, 并在平均前取绝对值将正面和负面贡献统一为重要性.

如图 10 所示, 在 MSTAR 数据集上使用了多种主要归因方法对多种模型进行实验分析. 图 10 中不同主要归因方法的显著性图之间存在差异, 从不同角度提供了不同的信息. 如 GGC 中的 Grad-CAM 针对的是最后一层卷积层, 因此上采样到输入图像空间后, 对识别重点刻画不够精细, 在图 10(b) 中表现较为明显. 这说明, 对于部分输入, VGG-16 (随机初始化) 在高层并未准确捕捉到目标位置信息, 导致其测试集识别率低于其他模型. 参考点的选取不同, 对显著性图也有明显的影响, 对比 IG 与 GradientSHAP, DeepLIFT 与 DeepSHAP, 将多张输入图像作为参考点, 可在显著图中凸显目标信息, 减弱随机的背景信息. 部分方法在跨模型下表现不佳, 可发现 GGC 并不适用于浅层模型、多种图

图 9 (网络版彩图) 输入为光学图像的部分卷积层特征图. (a) 输入图像; (b) ResNet-50 (随机初始化); (c) ResNet-50 (预训练); (d) ResNet-50 (ImageNet)

Figure 9 (Color online) Partial convolution feature maps. (a) Input image; (b) ResNet-50 (random initialization); (c) ResNet-50 (pre-trained); (d) ResNet-50 (ImageNet)

表 3 主要归因方法
Table 3 Primary attribution

Primary attribution	Description
GGC	Combine Guided-Backpropagation and Grad-CAM, consider positive contribution.
IG	Use integral to calculate importance, set the reference point to a single all-zero image.
LRP	Use the ε -rule rule to calculate the saliency map, set the parameter to $1E-9$.
DeepLIFT	Use the defined rules, set the reference point to a single all-zero image.
GradientSHAP	Use IG to calculate SHAP value, set the reference point to multiple input images.
DeepSHAP	Use DeepLIFT to calculate SHAP value, set the reference point to multiple input images.

像作为参考点的 SHAP 方法对目标的刻画不够精细等. 但大多数方法的显著性图之间较为相似. 说明模型的识别重点包括目标和背景干扰, 训练集的高识别率可能是由于去除了部分噪声, 如 VGG-16 (图 10(c)), 也有可能是利用了背景干扰如三层 CNN (图 10(a)) 和 ResNet-50 (图 10(d) 和 (e)).

如图 11 所示, 对三层 CNN 输入了 3 张识别错误的图像, 使用 IG 绘制不同类的显著性图, 借此可分析各类之间的相似程度. 图 11(a) 中各类的显著性图极为相似, 尚不明确模型是根据目标哪些部分区分不同类的, 只有目标细节和背景部分略有差异. 这说明在模型认知中, 不同类目标的雷达图像极为相似. 数据的相似性导致模型在主要依靠目标因果信息进行识别的同时, 会倾向于利用背景干扰信息辅助识别, 这使得模型鲁棒性降低. 图 11(b) 为考虑到各类显著性图强度而作的可视化, 不同类的置信度不同, 对应的显著性图强度也不一样, 图中显示模型错误识别结果更倾向于哪一类.

图 10 (网络版彩图) 显著性图可视化结果. (a) 三层 CNN (随机初始化); (b) VGG-16 (随机初始化); (c) VGG-16 (预训练); (d) ResNet-50 (随机初始化); (e) ResNet-50 (预训练)

Figure 10 (Color online) Visualization results of saliency map. (a) Three-layer CNN (random initialization); (b) VGG-16 (random initialization); (c) VGG-16 (pre-trained); (d) ResNet-50 (random initialization); (e) ResNet-50 (pre-trained)

图 11 的实验结果说明, 对于单通道 SAR 图像识别任务而言, 主要难点并不在于对信息的提取, 而在于不同类目标的相似性极高, 导致模型利用了相关干扰信息辅助识别. 为进一步说明该问题, 如图 12 所示, 输入 3 张光学图像, 使用 IG 绘制不同类的显著性图, 显示 SAR 目标识别模型对光学图像纹理的提取效果, 不同类结果均有一个相似的识别重点, 但差异性表现更为明显. 图 12(a) 和 (c) 的模型表现出了点状聚集特点, 图 12(c) 的 ResNet-50 (预训练) 对背景信息提取较多, 图 12(b) 的 VGG-16 (预训练) 仍保留了精细的纹理提取能力.

基于以上对于显著性图的分析, 可得出如下定性结论.

(1) 可信的主要归因方法应当符合人的主观认识, 并且能够客观反映模型, 在不同模型上均可得到良好的解释结果. 不同的主要归因方法得到的显著性图具有共性和差异性, 可得到相互印证的结论和不同方面的信息. 根据错误识别的显著性图, 可分析不同类之间识别重点的相似性, 并可诊断模型是

图 11 (网络版彩图) 基于主要归因方法的模型诊断. (a) 单张对比度调节; (b) 按行对比度调节

Figure 11 (Color online) Model diagnosis based on primary attribution. (a) Single contrast control; (b) horizontal contrast control

否正确提取了目标信息.

(2) 在 SAR 数据集中, 由于目标和背景干扰均为散射点, 目标和背景、不同类目标彼此之间具有一定的相似性, 导致在背景干扰较强的情况下, 深度学习模型会根据目标和背景干扰之间的相关性辅助识别. 因此在数据集较小的情况下, 模型高识别率可能是由于模型利用了背景干扰与目标的相关性, 导致在实际任务中模型鲁棒性较差. 模型鲁棒性可表述为在显著性图上对背景干扰信息的去除能力, 具有鲁棒性的模型应当聚焦于与目标具有因果性的信息, 而不是强相关性的干扰信息.

4.3.2 基于主要归因方法的模型鲁棒性评估

本小节采用了表 4 的指标对主要归因方法与深度学习黑盒模型进行评估, 其中 RI 代指随机初始化, PT 代指预训练. 根据表 4 计算得到表 5 的结果, 通过对多张图像进行多次采样, 确保指标的稳定性. 表 5 失真度和最大灵敏度结果显示, IG 在各个 SAR 目标识别模型中表现出了较好的准确性和鲁棒性, 因此基于 IG 在训练集上计算得到 IoU.

表 5 的 IoU 结果说明 VGG-16 (预训练) 在鲁棒性上是优于其他模型的, 这与表 2 中测试集识别率和基于图 10 和 12 的显著性图分析结果相吻合. VGG-16 (预训练) 测试集识别率最高, 并且显著性图中对于目标特征刻画较为精细, 背景干扰明显少于其他模型. 这说明模型学习到了更为精确的特征表达方式, 有利于提高在实际任务中的鲁棒性.

对 ResNet-50 (预训练) 而言, 测试集识别率高, 但 IoU 指标低, 这说明模型更多地利用了背景干扰信息辅助识别, 图 10 和 12 的显著性图结果也反应模型过拟合背景干扰信息, 由于测试集仅是俯仰

图 12 (网络版彩图) 基于主要归因方法的模型分析. (a) 三层 CNN (随机初始化); (b) VGG-16 (预训练); (c) ResNet-50 (预训练)

Figure 12 (Color online) Model analysis based on primary attribution. (a) Three-layer CNN (random initialization); (b) VGG-16 (pre-trained); (c) ResNet-50 (pre-trained)

角变化, 因此背景干扰的相关性有助于提高测试集识别率和增加图 8 中预训练模型的特征可分性. 但小样本情况下对背景干扰的过拟合并不是实际任务中需要的特性, 这会造成在大样本情况下鲁棒性的下降.

从可解释性角度评估模型鲁棒性, 可有效分析模型因果信息提取能力, 提供不同于测试集识别率评估的新角度, 为实际任务决策提供了更多可信解释.

5 结论与展望

综上所述, 本文首次从可解释性方面分析了雷达图像领域中的深度学习模型, 较为系统地提出了

表 4 评估指标

Table 4 Evaluation measures

Evaluation measures	Description
Infidelity ^[43]	The accuracy of attribution algorithms
Max-Sensitivity ^[43]	Robustness of attribution algorithms
IoU	Robustness of approximate measurement model

表 5 评估指标结果

Table 5 Evaluation measures results

Evaluation measure	Attribution algorithm	Three-layer CNN (RI)	VGG-16 (RI)	VGG-16 (PT)	ResNet-50 (RI)	ResNet-50 (PT)
Infidelity	GGC	0.5274	0.8909	0.8909	0.6398	0.9053
	IG	0.3603	0.6128	0.6005	0.4453	0.6348
	LRP	0.5260	0.8989	0.8906	0.6405	0.9130
	DeepLIFT	0.5193	0.8899	0.8801	0.6171	0.8719
	GradientSHAP	0.3860	0.7592	0.7092	0.5011	0.6723
	DeepSHAP	0.5265	0.8902	0.8902	0.6397	0.9035
Max-sensitivity	GGC	1.327	7532	4.327	4.924	199.7
	IG	0.3123	0.2901	0.1845	0.4547	0.7620
	LRP	255.2	268.6	52.71	326.5	8451
	DeepLIFT	0.5545	6.089	0.5338	0.8743	2.297
	GradientSHAP	21.43	56.49	4211	12094	75.78
	DeepSHAP	0.4742	5.679	0.7341	0.6596	7.860
IoU	IG	0.3653	0.4615	0.4639	0.3970	0.3576

雷达图像深度学习黑盒模型分析思路,并在MSTAR数据集上进行了较充分的实验验证分析,主要完成了以下两点工作。

(1) 在模型内部对模型迁移机制的研究,分析了雷达和光学图像的差异性影响,发现迁移机制的成功在于底层纹理提取机制具有较强的一般性,迁移后可提取雷达图像的有用信息,但数据的差异性导致模型冗余度增加。并且雷达图像的特征提取机制并不唯一,直接训练的点状分布卷积核和迁移的光学纹理卷积核同样有效。不同源的数据包含不同纹理特征,可以提高模型对不同类型目标的区分度。

(2) 在模型整体上对模型认知机理的研究,分析了雷达目标之间、目标和背景之间的相似性影响,发现雷达图像的相似性导致模型根据目标和背景干扰的相关性辅助识别。基于主要归因方法分析诊断了模型的认知机理,并使用了客观指标分析主要归因方法在目标识别模型上的准确性和鲁棒性,为主要归因方法应用于雷达图像领域深度学习黑盒模型分析提供了基础。从可解释性角度评估模型鲁棒性,反映了模型对因果信息的表征能力,为SAR目标识别模型鲁棒性评估提供不同思路。

未来可从以下几个方面研究深度学习技术在雷达图像领域的进一步应用。

(1) 对模型迁移机制的研究。光学预训练模型有助于提高性能,但带来的冗余度需要进一步研究。目前,已有研究通过模型压缩寻找一个稀疏子网络^[11],或利用归因方法Grad-CAM评估模型神经元的因果重要性^[51]进行压缩,在雷达领域也有关于SAR目标识别模型压缩的相关报道^[52],但未与迁移学习的冗余度相结合得到一个轻量化模型。依据本文结论,可在文献[51]的基础上利用层归因方法

评估神经元重要性, 只保留光学预训练模型的部分卷积核, 从而减少模型参数, 提高 SAR 目标识别模型效率.

(2) 对模型底层特征提取机制的研究. 卷积核能有效提取光学目标的尺度和纹理信息, 而雷达图像相比于光学图像, 其纹理特征并不明显, 且背景干扰信息显著, 需结合雷达领域知识设计与散射点纹理特征相对应的卷积核. 基于可变形卷积网络^[53] (deformable convolutional networks) 可自适应学习与数据对应的卷积核结构, 或利用雷达图像领域物理知识, 设计一个表征目标散射点结构信息的特征提取器, 使模型具有清晰的规则约束.

(3) 对模型鲁棒性的研究. 由于雷达图像背景干扰信息显著, 并且高价值目标的雷达数据较少, 模型会借助非因果信息进行识别. 文献 [54] 利用自编码器构建深度学习模型消除雷达图像背景干扰信息. 文献 [55] 从对抗样本角度分析模型鲁棒性, 将模型学习的特征分为鲁棒性特征 (符合人的主观认识) 和非鲁棒性特征 (人无法理解的特征), 对抗样本寻找到了人无法理解但对模型识别十分重要的非鲁棒性特征. 可基于文献 [44, 55], 研究利用对抗攻击训练模型, 剔除非鲁棒性特征, 提高模型的可解释性和鲁棒性. 此外, 本文实验结果显示目标因果信息包含于模型的一阶和高阶偏导数级数中, 背景相关信息主要包含于模型的一阶偏导数中. 这说明模型在学习这两种不同信息时已经产生差异, 后续将研究基于二阶导数的元学习算法, 利用模型的高阶偏导数增强模型鲁棒性. 另一种研究思路是利用背景环境的信息, 从而根据目标与环境的差异进行目标识别.

(4) 对模型可解释性的研究. 数据驱动的深度学习模型学习到了超出人现有认知能力的相关知识, 但很可能包含了非因果信息, 需要进一步研究模型内部知识. 对于模型的内部分析, 文献 [17] 提出的可解释性卷积核包含明确的语义信息, 如顶层卷积核通常包含动物的头部信息, 借助此可进一步分析目标不同部件对于模型的影响, 为决策提供更多信息. 主要归因方法中正负贡献分析、参考点的选择等方面也有待进一步分析.

尽管模型的所有参数和结构是已知的, 但是模型的认知机理、工作机制仍然是一个无法理解的黑盒. 因此本文对黑盒模型进行了初步的研究探索, 主要关注在雷达图像中深度学习模型怎么提取特征, 特征提取的结果是什么, 怎么评价特征提取的好坏, 但距离揭开黑盒模型仍然很远. 下一步将主要研究如何利用可解释性分析结论改进模型, 在逐步揭开黑盒模型的过程中, 利用可解释性好的数学、物理等理论与方法, 使得黑盒模型中一部分具备自解释性, 由此逐渐实现半白盒模型应用于雷达信息处理, 增强其应用的可信度.

参考文献

- 1 Wang J, Zheng T, Lei P, et al. Study on deep learning in radar (in Chinese). J Radars, 2018, 7: 395–411 [王俊, 郑彤, 雷鹏, 等. 深度学习在雷达中的研究综述. 雷达学报, 2018, 7: 395–411]
- 2 Pan Z X, An Q Z, Zhang B C. Progress of deep learning-based target recognition in radar images (in Chinese). Sci Sin Inform, 2019, 49: 1626–1639 [潘宗序, 安全智, 张冰尘. 基于深度学习的雷达图像目标识别研究进展. 中国科学: 信息科学, 2019, 49: 1626–1639]
- 3 Keydel E R, Lee S W, Moore J T. MSTAR extended operating conditions: a tutorial. In: Proceedings of SPIE, 1996. 228–242
- 4 Fjelland R. Why general artificial intelligence will not be realized. Humanit Soc Sci Commun, 2020, 7: 10
- 5 Zhang Q S, Wang W G, Zhu S C. Examining CNN representations with respect to dataset bias. 2017. ArXiv:1710.10577
- 6 Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell, 2019, 267: 1–38
- 7 Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. 2020. ArXiv:2006.11371
- 8 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2014. 818–833

- 9 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. 2013. ArXiv:1312.6034
- 10 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. 2013. ArXiv:1312.6199
- 11 Frankle J, Carbin M. The lottery ticket hypothesis: finding sparse, trainable neural networks. 2018. ArXiv:1803.03635
- 12 Hu Z T, Ma X Z, Liu Z Z, et al. Harnessing deep neural networks with logic rules. 2016. ArXiv:1603.06318
- 13 Park D H, Hendricks L A, Akata Z, et al. Multimodal explanations: justifying decisions and pointing to the evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 8779–8788
- 14 Wang H, Yeung D Y. Towards Bayesian deep learning: a framework and some existing methods. *IEEE Trans Knowl Data Eng*, 2016, 28: 3395–3408
- 15 Liu H D, Gu X F, Samaras D. A two-step computation of the exact GAN wasserstein distance. In: Proceedings of International Conference on Machine Learning, 2018. 3165–3174
- 16 Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. 2017. ArXiv:1710.09829
- 17 Zhang Q S, Wu Y N, Zhu S C. Interpretable convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 8827–8836
- 18 Guo W W, Zhang Z H, Yu W X, et al. Perspective on explainable SAR target recognition (in Chinese). *J Radar*, 2020, 9: 462–476 [郭炜炜, 张增辉, 郁文贤, 等. SAR 图像目标识别的可解释性问题探讨. *雷达学报*, 2020, 9: 462–476]
- 19 Kokhlikyan N, Miglani V, Martin M, et al. Captum: a unified and generic model interpretability library for PyTorch. 2020. ArXiv:2009.07896
- 20 Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net. 2014. ArXiv:1412.6806
- 21 Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 618–626
- 22 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of International Conference on Machine Learning, 2017. 3145–3153
- 23 Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One*, 2015, 10: e0130140
- 24 Binder A, Montavon G, Lapuschkin S, et al. Layer-wise relevance propagation for neural networks with local renormalization layers. In: Proceedings of International Conference on Artificial Neural Networks. Cham: Springer, 2016. 63–71
- 25 Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of International Conference on Machine Learning, 2017. 3319–3328
- 26 Ancona M, Ceolini E, Öztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks. 2017. ArXiv:1711.06104
- 27 Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu.com, 2019. 154–162. <https://christophm.github.io/interpretable-ml-book/>
- 28 Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*, 2019, 20: 1–81
- 29 Ribeiro M T, Singh S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. 1135–1144
- 30 Ribeiro M T, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. 2016. ArXiv:1606.05386
- 31 Smilkov D, Thorat N, Kim B, et al. Smoothgrad: removing noise by adding noise. 2017. ArXiv:1706.03825
- 32 Adebayo J, Gilmer J, Goodfellow I, et al. Local explanation methods for deep neural networks lack sensitivity to parameter values. 2018. ArXiv:1810.03307
- 33 Shapley L S. A value for n-person games. In: *Contributions to the Theory of Games II*. Princeton: Princeton University Press, 1953. 307–317
- 34 Young H P. Cost Allocation: Methods, Principles, Applications. Amsterdam: North Holland Publishing Company, 1985. 69–94
- 35 Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*, 2014, 41: 647–665
- 36 Castro J, Gómez D, Tejada J. Polynomial calculation of the Shapley value based on sampling. *Comput Oper Res*, 2009, 36: 1726–1730
- 37 Lundberg S M, Lee S I. A unified approach to interpreting model predictions. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 4765–4774

- 38 Dhamdhere K, Sundararajan M, Yan Q. How important is a neuron? 2018. ArXiv:1805.12233
- 39 Leino K, Sen S, Datta A, et al. Influence-directed explanations for deep convolutional networks. In: Proceedings of IEEE International Test Conference (ITC), 2018. 1–8
- 40 Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018. 839–847
- 41 van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605
- 42 McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. 2018. ArXiv:1802.03426
- 43 Yeh C K, Hsieh C Y, Suggala A S, et al. On the (in) fidelity and sensitivity for explanations. 2019. ArXiv:1901.09392
- 44 Etmann C, Lunz S, Maass P, et al. On the connection between adversarial robustness and saliency map interpretability. 2019. ArXiv:1905.04172
- 45 Seo J, Choe J, Koo J, et al. Noise-adding methods of saliency map as series of higher order partial derivative. 2018. ArXiv:1806.03000
- 46 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 47 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
- 48 Huang Z L, Pan Z X, Lei B. What, where, and how to transfer in SAR target recognition based on deep CNNs. IEEE Trans Geosci Remote Sens, 2020, 58: 2324–2336
- 49 Wen Y D, Zhang K P, Li Z F, et al. A discriminative feature learning approach for deep face recognition. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016. 499–515
- 50 Marmanis D, Yao W, Adam F, et al. Artificial generation of big data for improving image classification: a generative adversarial network approach on SAR data. In: Proceedings of International Conference on Big Data from Space, 2017
- 51 Zhang B, Yang P B, Sang J T, et al. Convolution network pruning based on the evaluation of the importance of characteristic attributions (in Chinese). Sci Sin Inform, 2021, 51: 13–26 [张彪, 杨朋波, 桑基韬, 等. 基于特征归因重要度评估的卷积网络剪枝. 中国科学: 信息科学, 2021, 51: 13–26]
- 52 Min R, Lan H, Cao Z, et al. A gradually distilled CNN for SAR target recognition. IEEE Access, 2019, 7: 42190–42200
- 53 Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 764–773
- 54 Li C, Du L, Deng S, et al. Point-wise discriminative auto-encoder with application on robust radar automatic target recognition. Signal Process, 2020, 169: 107385
- 55 Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features. 2019. ArXiv:1905.02175

Research and exploration on the interpretability of deep learning model in radar image

Weijie LI, Wei YANG*, Yongxiang LIU & Xiang LI

College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China

* Corresponding author. E-mail: yw850716@sina.com

Abstract Although deep learning technology has been partially applied in radar image processing, the lack of interpretability analysis and comprehensive performance evaluation of black-box models limits the application performance, credibility, and universality of this technology in the radar image field. Starting from the interpretability, this paper proposes the mechanism analysis idea of deep learning black-box model in the radar image field and carries out experimental verification on the open-source MSTAR radar image dataset. It analyzes the transfer and cognitive mechanisms of the deep learning model and draws significant findings on transfer learning, attribution method application, and model robustness evaluation, filling gaps in the existing literature.

Keywords deep learning, radar image, interpretability, transfer learning, attribution algorithm, robustness