计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):228−234 [doi: 10.15888/j.cnki.csa.006937]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的铁路图像场景分类优化研究①

赵　冰[1], 李　平[2], 代明睿[2], 马小宁[2]

[1](中国铁道科学研究院 研究生部, 北京 100081)
[2](中国铁道科学研究院 铁路大数据研究与应用创新中心, 北京 100081)
通讯作者: 赵　冰, E-mail: 14121378@bjtu.edu.cn

摘　要: 铁路检测、监测领域产生海量的图像数据, 基于图像场景进行分类对图像后续分析、管理具有重要价值. 本文提出一种结合深度卷积神经网络 DCNN (Deep Convolutional Neural Networks) 与梯度类激活映射 Grad-CAM (Grad Class Activation Mapping) 的可视化场景分类模型, DCNN 在铁路场景分类图像数据集进行迁移学习, 实现特征提取, Grad-CAM 根据梯度全局平均计算权重实现对类别的加权热力图及激活分数计算, 提升分类模型可解释性. 实验中对比了不同的 DCNN 网络结构对铁路图像场景分类任务性能影响, 对场景分类模型实现可视化解释, 基于可视化模型提出了通过降低数据集内部偏差提升模型分类能力的优化流程, 验证了深度学习技术对于图像场景分类任务的有效性.

关键词: 深度学习; 铁路图像; 场景分类; 可视化; 迁移学习

引用格式: 赵冰,李平,代明睿,马小宁.基于深度学习的铁路图像场景分类优化研究.计算机系统应用,2019,28(6):228–234. http://www.c-s-a.org.cn/1003-3254/6937.html

Research on Optimization Method of Railway Image Scene Classification Based on Deep Learning Method

ZHAO Bing[1], LI Ping[2], DAI Ming-Rui[2], MA Xiao-Ning[2]

[1](Department of Postgraduates, China Academy of Railway Sciences, Beijing 100081, China)
[2](Railway Big Data Research Center, China Academy of Railway Sciences, Beijing 100081, China)

Abstract: The field of railway detection and monitoring generates massive image data, image scene classification is of great value for subsequent analysis and management. In this study, a visual scene classification model that combines Deep Convolutional Neural Networks (DCNN) and Grad Class Activation Mapping (Grad-CAM) is proposed, DCNN extract feature of railway scene classification image dataset by transfer learning method, Grad-CAM improves the interpretability of the classification model by calculating the weighted thermogram and activation scores of the categories. In the experiment, the effects of different DCNN structures on the performance of railway image scene classification tasks are compared, and visual interpretation of scene classification model is realized. At the same time, based on visualization method, an optimization process is proposed to improve model classification ability by reducing internal deviation of dataset, which verifies the effectiveness of the deep learning technology for image scene classification task.

Key words: deep learning; railway image; scene classification; visualization; transfer learnin

铁路检修及监控任务产生大量的图像数据, 图像　　后续分析一般与图像拍摄场景相关联, 但目前图像储

① 基金项目: 铁科院院基金重大课题 (2017YJ005)
Foundation item: Major Project of China Academy of Railway Sciences (2017YJ005)
收稿时间: 2018-12-10; 修改时间: 2018-12-29; 采用时间: 2019-01-10; csa 在线出版时间: 2019-05-25

存后缺乏科学有效图像场景分类方法, 限制了图像价值的分析与挖掘. 由于铁路领域缺乏专业大规模图像数据集, 与自然图像相比, 图像包含大量前景语义, 铁路领域因素明显, 同时复杂光照条件、恶劣环境因素、设备素质局限均影响成像效果, 增加了铁路图像场景分类任务的难度. 快速准确的图加了铁路图像场景分类任务的难度. 快速准确的图像场景自动分类方法有助于图像分析研究工作, 可为铁路部件检测、安保监控、周界防护等研究工作提供图像分类预处理流程, 具有切实研究应用价值.

目前图像场景分类方法可分为两类: 基于手动特征提取方法与基于深度学习的方法. 基于手动特征提取的分类方法在几十年的发展中已趋于完善, Harris 角点检测子[2]、DoG 算子[3](Difference of Gaussian) 等兴趣点检测算法选择局部明显特征, 在较小的计算开销下能够获得一定的空间几何不变性; 方向梯度直方图 HOG[4](Histogram of Oriented Gradient)、尺度不变特征转换 SIFT[5](Scale Invariant Feature Transform)、局部二值模式 LBP[6](Local Binary Pattern) 等方法采取密集的特征提取方式, 结合支持向量机 SVM[7](Support Vector Machine) 获得了更优的特征提取效果. 基于深度学习方法的 CNN 网络可对数据进行自动特征提取, 克服了传统提取特征方法需要手动设计特征提取算子的弊端, 随着 2012 年 ImageNet 大规模视觉识别挑战赛上 AlexNet 超越传统方法, 在图像分类任务上取得优异成绩后, 后续 VGG、GoogLeNet、InceptionNet、ResNet 等 CNN 网络模型的提出, 使得基于 CNN 的深度学习技术成为图像分类任务的主流方法[8]. 尽管表现出色, 但 CNN 模型的可解释性问题一直为人诟病, 在对 CNN 进行可视化的解释的方面, Zeiler 等人[9] 提出建立反向卷积神经网络, 产生高分辨率和可理解的特征可视化图像; 类激活映射 CAM (Class Activation Mapping)[10]通过改变 CNN 的网络结构并对数据集进行重新训练, 获得类别定位激活图, 观察 CNN 的感兴趣区域; 梯度类激活映射 Grad-CAM (Grad Class Activation Mapping) 可对训练好的分类网络模型输出计算梯度全局平均权重, 减少了模型重构及训练的时间.

本文提出了将 DCNN 应用于铁路图像进行特征提取实现铁路图像场景分类, 所建立的网络结构级联 Grad-CAM 可视化层实现模型原理可解释性. 本文对比了不同的 DCNN 网络在铁路场景图像下迁移学习的能力, 同时提出了通过可视化方法降低数据集内部偏差提升模型分类能力的优化流程, 结果显示本文提出方法可有效降低数据集偏差, 提升模型分类能力.

1 研究方法

为实现在小样本量下的铁路图像场景分类任务, 本文基于迁移学习思想, 将在 ImageNet 数据集上预训练好的网络模型进行迁移调参训练, 比较了现有网络及本文改进网络在铁路图像场景分类任务下的性能; 为实现 CNN 分类模型可解释性, 本文基于 Grad-CAM 思想, 在经过下采样后的最后一层特征映射图上, 对梯度权重添加 ReLU 层, 实现图像不同区域对分类类别激活的热力图解释. 全文研究流程如图 1 所示.

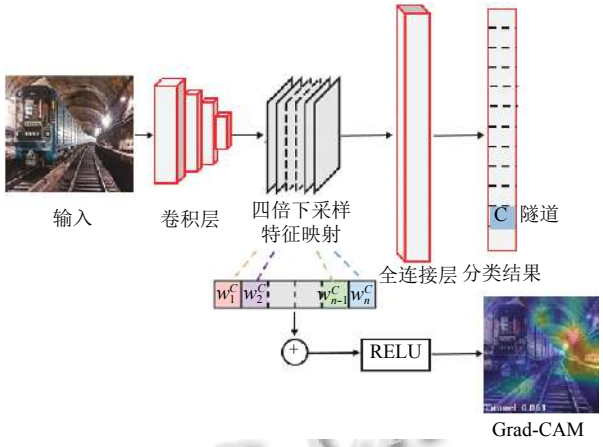


图 1 铁路图像场景分类及可视化分析流程

1.1 迁移学习

为避免 CNN 模型在小型场景数据集训练出现难以收敛的问题, 本文将在 ImageNet 数据集上预训练好的网络模型迁移至铁路场景数据集进行训练, 表 1 展示 ResNet50 和 ResNet101 的模型结构, 图片进入模型后裁剪成固定分辨率, 经过 CNN 完成特征提取工作, 结合本文特定问题, 提出改进后 ResNet50 网络模型如下:

1) 在 conv1、conv2、conv3、conv4 层后分别连 Dropout 层抑制过拟合, Dropout 率设定为 0.2; 2) 全连接层修改为 12 维, 接 softmax 层完成对各类别概率计算; 3) 固定前端所有层参数, 仅对最后一层全连接层进行调参训练.

以 ResnetV1-50 模型为例, 如图 2 所示, 图片输入尺寸为 224×224 像素, 在第一阶段中, 图像经过卷积层 1, 卷积核为 7×7 像素, 步长为 2 像素, 维数为 64, 输出

尺寸为 112×112 像素, 进而连接批量归一化层及非线性变换层以加快收敛, 最后经过卷积核为 3×3 像素, 步长为 2 像素的最大池化层; 第二阶段, 图像经过一个层数为 3 的残差块, 标准残差块的卷积核为 1×1 或 3×3 像素, 输出的图像尺寸为 56×56 像素; 第三阶段图像经过一个层数为 4 的残差块, 输出尺寸为 28×28 像素; 第四阶段图像经过一个层数为 6 的残差块, 输出尺寸为 14×14 像素; 第五阶段图像经过一个层数为 3 的残差块, 输出尺寸为 7×7 像素; 最终经过平均池化层、12 维的全连接层及归一化指数层, 求得图像分类概率.

1.2 CNN 可视化模型

(1) 梯度反向传播 (guided backpropagation)[11]:

对 CNN 各层输入和梯度均大于 0 的所对应的梯度进行反向传播, 可在一定程度上展示该层提取到的全部特征, 相邻步梯度定义为:

$$R_i^l = (R_i^{l+1} > 0) \cdot R_i^{l+1} \tag{1}$$

其中, Ri 为迭代过程中的梯度, 根据反向传播可由 l+1 步的梯度求解出 l 步的梯度.

表 1 ResNet 网络结构

Layer name	Output size	50-layer	101-layer
conv1	112×112	7×7, 64, stride2	
		3×3 maxpool, stride2	
conv2_x	56×56	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4
conv4_x	14×14	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23
conv5_x	7×7	[1×1, 256; 3×3, 256; 1×1, 1024] ×3	[1×1, 256; 3×3, 256; 1×1, 1024] ×3
	1×1	average pool, 12-d fc, softmax	

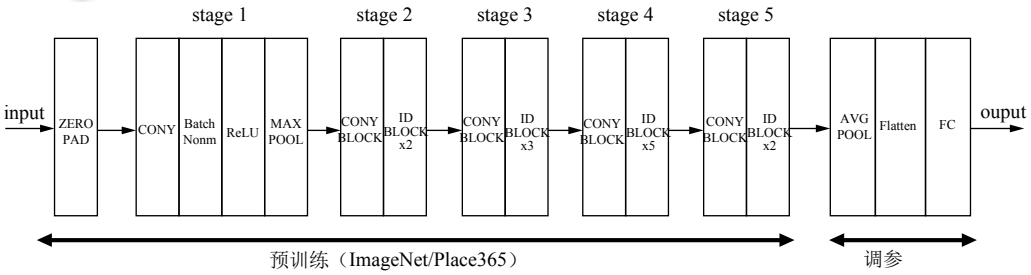


图 2 ResNet50 网络迁移学习流程

(2) 类激活映射 (Class Activation Mapping)[12]:

通过多次卷积和池化后, CNN 最后一层卷积层包含了丰富的空间和语义信息, 引入全局平均池化替换全连接层, 保留下丰富的语音信息, 对每一个特定类别 c, 类激活映射可显示其类激活映射, 如:

$$M_c(x,y) = \sum_k \varpi_k^c f_k(x,y) \tag{2}$$

$$L_{CAM}^C = \sum_k \varpi_k^c A_{ij}^k \tag{3}$$

其中, MC 指针对类别 C 的类别激活映射, $f_k(x,y)$表最后一层卷积层上单元 k 的位置, L_{CAM}^C指利用 CA 方法生成的激活值.

(3) 渐变类激活映射 (Grad Class Activation Mapping, Grad-CAM)[13]:

利用梯度全局平均来计算权重可在不改变原网络结构的情况下实现对类别的加权热力图及激活分数, 如:

$$L_{Grad-CAM}^c = \mathrm{Re}LU\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

$$S^c = \frac{1}{Z}\sum_i\sum_j\sum_k \varpi_k^c A_{ij}^k \tag{5}$$

其中, $\sum_k \alpha_k^c A^k$代表线性集合, 对其进行 RELU 激活操作可得针对特定类别 c, Grad-CAM 方法生成的激活值 $L_{Grad-CAM}^c$.

2 铁路场景分类数据集

ImageNet[14]、SUN [15]、Places[16]等大型数据集的提出推动了深度学习的发展, 但这些数据集仅关注自

然图像, 目前没有专门用于铁路目标或场景分类的铁路图像 (如包含接触网, 涵洞, 机车, 线路等铁路专有目标) 数据集. 为了弥补缺乏数据集对深层学习方法的制约, 本文制作了一个名为 Railway12 的铁路场景图像数据集, , Railway12 包含 12 种典型场景 (隧道、候车大厅、购票大厅、火车站广场、接触网、维修车间、站台、货场、车厢、铁路道口、铁路桥梁、防护网), 数据集总图片量为 1.2 万幅图像, 每个类别包含 1000 张图像. 图 3 显示了 Railway12 数据集的一些图像样本.

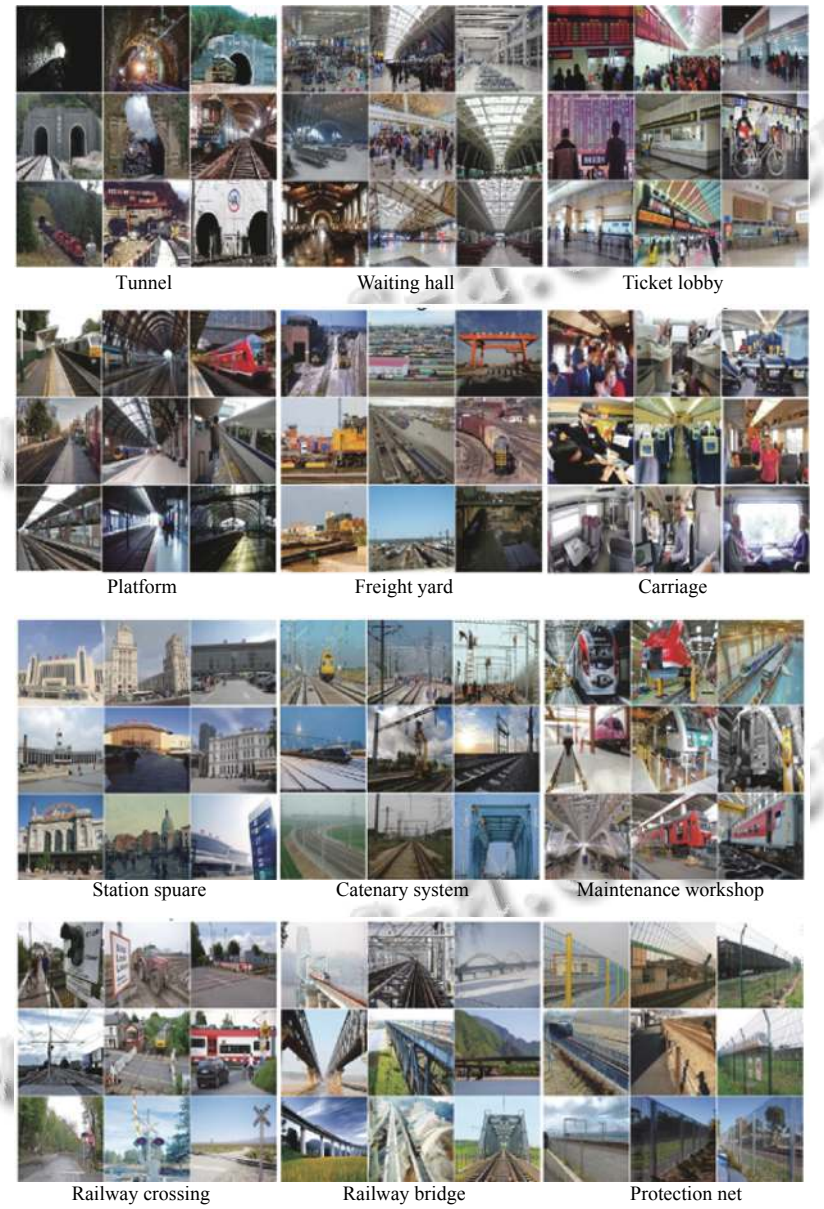


图 3 Railway12 数据集图像样例

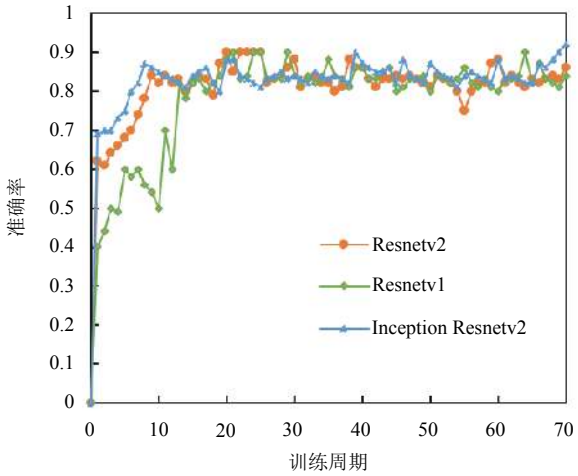
3 实验结果分析

3.1 场景分类结果

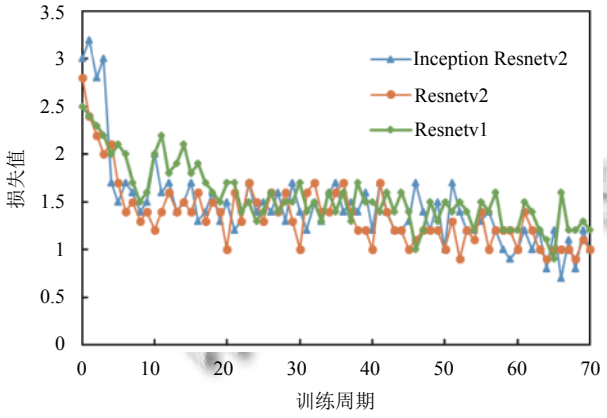
本节对深层 CNN 的场景分类能力进行了评估. 选择三种典型的 TF-Slim CNN 模型 (ResNetV1, ResNetV2, Inception-ResNet-v2) 进行训练, 这些模型已经在 ILSVRC-2012-CLS 图像分类数据集预先训练过, 以具有良好的并行处理能力. 实验在数据集按 7: 2: 1 的比例划分为训练集、验证集、测试集. 出于后续工程化考虑, 实验在 linux 环境下基于 Tensorflow 深度学习框架搭建, 计算机配置为: Inter Core i7, 显卡为

1080Ti, 参数设置如下: 学习率 (learning_rate)=0.0001, 训练批次尺寸 (batch_size)=32, 训练周期 (num_epochs)=75–85. 除 logits 层之外, 所有网络层均固定参数. 我们在表 1 中提供了训练集的 top-3 的分类精度, 以及测试集的 top-1 和 top-3 的分类精度.

实验结果表明, 将在 ImageNet 预训练的 CNN 模型迁移到铁路场景分类任务是可行的, 模型分类精度优异, 考虑到铁路行业庞大的图像数据量, 模型具有实际的应用价值. 实验表明, 网络结构越深, 提取特征模型的能力越强. 模型的精度和损失历史如图 4 所示.



(a) 模型训练过程准确率演化



(b) 模型训练过程损失值演化

图 4 模型训练准确率及损失历史

3.2 模型可视化分析

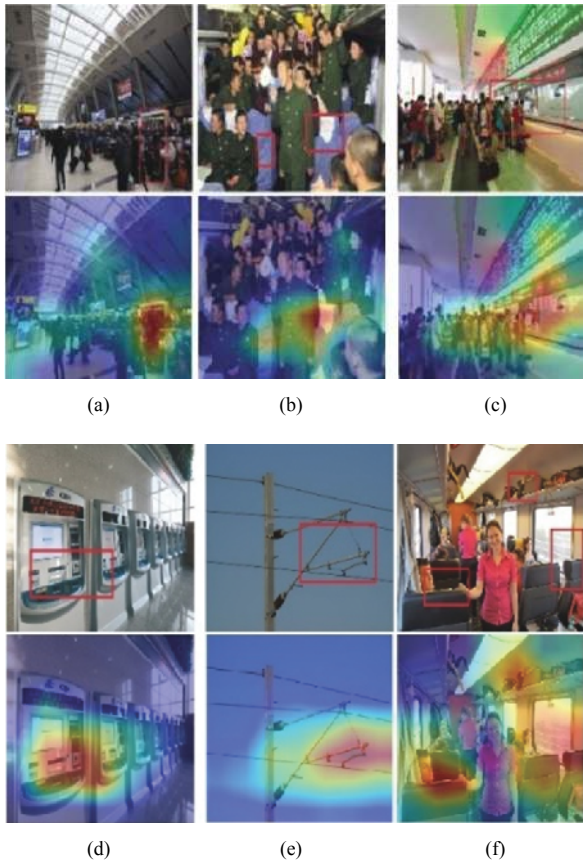
3.2.1 模型激活区域分析

利用 Grad-CAM 方法将 CNN 最后一层卷积层的梯度信息生成类激活图, 热力图可以显示模型对不同类别感兴趣的区域, 根据这种方法, 我们可以对 CNN 模型的训练结果进行直观解释.

本文发现丰富的前景语义不是造成分类错误的主要原因 (如图 5(a) 和 (b), 在图像包含大量的前景语义情况下, 模型仍根据背景信息实现了正确的分类.

表 2 不同网络结构的分类准确率

Accuracy (%)	Train (Top3)	Test (Top1)	Test (Top3)
ResNet V1	95.2	79.8	94.6
ResNet V2	95.8	80.3	94.9
Inception-ResNet-v2	96.7	82.6	95.3



(a) (b) (c)

(d) (e) (f)

图 5 CNN 对不同图像类型感兴趣的区域: (a) 和 (b) 为具有丰富前景语义的图像, (c) 和 (d) 为具有很大类内差异的图像, (e) 和 (f) 具有不同场景复杂度的图像

当类别内部图片差异性较大时, 模型在训练中会形成不同的关注模式, 对于售票大厅类别, 一部分图片由售票窗口组成 (如图 5(c)), 其他图片由自动售票机 (如图 5(d)) 组成. 网络仍然可以进行正确判断.

当构成场景的对象较少时, 场景比较直观, 此时神经网络的判断策略类似于目标分类, 根据图像中出现的典型目标判断其归属类别, 如图 5(e) 所示; 当图像场景更加复杂时, 构成场景的对象是随机而大量的, 它将场景理解为前景和背景语义的组合, 作为如图 5(f) 所

示, 座椅、行李、窗户和乘客区域均对分类有贡献.

一般情况下, 我们只能从 CNN 模型输出中获知 top-n 预测类别和概率值. 但如图 6 所示, 利用 Grad-CAM 方法可以对同一张图片的不同预测结果展示模型感兴趣的区域. 尽管模型对这张图片做出了错误的判断, 但仍可以帮助我们分析模型做出错误分类的依据, 为看似不合理的分类结果提供合理的解释.

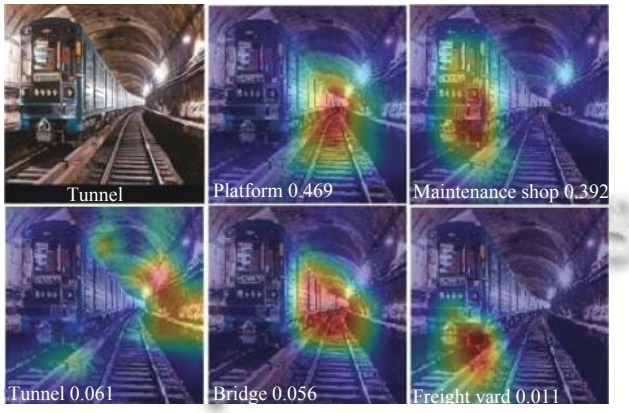


图 6 模型对同一输入图片的 top5 预测结果及 Grad-CAM 可视化展示

3.2.2 模型激活特征分析

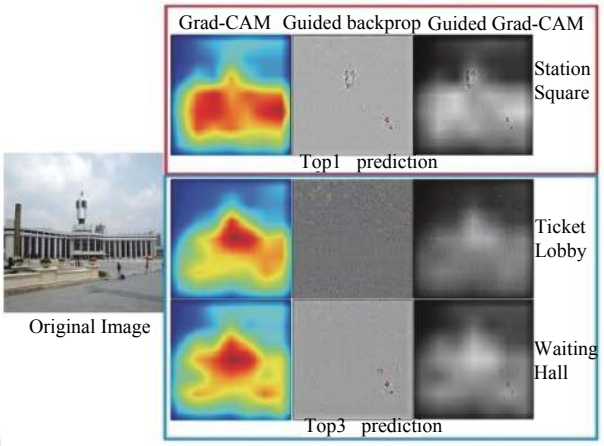
在 Grad-CAM 方法的基础上, 结合梯度信息可以揭示在 CNN 的决策过程中激活了哪些特征, 对 CNN 的运行机制做出深入的解释.

如图 7(a) 所示, 网络在 top1 预测中进行正确预测, 网络提取建筑尖顶作为显著特征, 而在 top3 中的其他预测中, 模型提取的特征为窗户和行人, 这些特征导致了错误的预测; 但在图 7(b) 中, 模型在 top1 预测中做出了错误的预测, 但当模型做出让人疑惑的决策时, Grad-CAM 和 Guided Grad-CAM 可以解释模型受到了怎样的干扰.

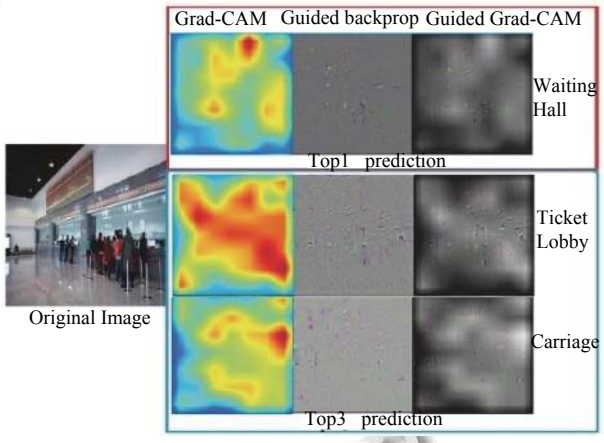
3.3 降低数据集偏差优化模型分类性能

本部分我们分别计算测试集中 12 类别图片的错误率, 结果见表 3. 选定 ResNet V1 模型作为基准, 分析数据集对于模型分类准确率的影响模式. 我们发现维修车间, 车厢, 桥梁场景为分类准确率最高的三个场景, 而隧道, 候车大厅, 车站为分类准确率最低的三个场景.

分类准确率较高的类别均体现类别内场景单一, 可识别目标较明显的特点; 分类准确率较低的类别体现出类内场景丰富的特点, 例如不同国家的火车站广场建筑风格差异巨大, 同时由于可能出现类别间相似性较大的问题, 导致某些类别的分类准确率较低.



(a) 模型top1预测正确



(b) 模型top1预测错误

图 7 结合多种可视化方法展示对于预测对应激活特征

表 3 测试集不同类别图片的准确率

Top3 Accuracy %	ResNet V1	ResNet V2	Inception-ResNet-v2
隧道	**86.8**	**87.2**	**9.1**
候车大厅	**88.2**	**87.8**	**88.5**
售票大厅	88.7	89.3	89.7
站前广场	**87.4**	**87.6**	**88.2**
接触网	92.1	93.4	93.6
维修车间	95.3	95.4	96.2
站台	93.6	94.3	94.5
货场	84.2	87.5	86.7
车厢	94.8	94.2	94.6
道口	93.6	94.6	95.3
铁路桥梁	95.2	95.4	95.6
防护网	92.4	94.8	96.3

数据集偏差会对分类准确率造成严重的影响, 当数据集出现各类别图片数据量相差较大时, 网络模型对各类图片的分类准确率会出现明显波动, 可以考虑使用数据增强技术进行弥补; 应注意训练集中同一场

景类别内各种类型图片的比例, 如果某种类型数量占比过大, 其他类型占比过小, 在这样有偏差的数据集训练得到的模型其泛化能力较弱, 容易产生有偏差和定势的模型.

4 结束语

在本文中, 我们建立了铁路专用场景分析数据集 Railway12, 基于迁移学习的思想, 将主流深度网络模型迁移到场景分类任务中, 获得能够胜任于铁路场景分类任务的高准确率模型. 利用 Grad-CAM 等可视化方法, 直观的解释网络的工作原理, 从网络感兴趣区域和分类激活特征两个层面分析模型在场景分类任务中的典型判断模式. 最终, 利用可视化方法分析数据集存在的偏差, 通过改进数据集提高模型的场景分类准确率. 未来我们计划对数据集建立准确的类别字典, 规范数据集类别内各子类图片比例, 完善数据集结构, 提高数据集统计学分布意义, 同时利用可视化手段指导网络模型的修改, 提出更加适用于铁路场景分类的网络模型.

参考文献

1 张萼辉. 铁路货车装载状态图像智能识别系统. 中国铁路, 2017, (9): 113–116.

2 Harris C, Stephens M. A combined corner and edge detector. Proceedings of the 4th Alvey Vision Conference. Manchester, UK. 1988. 147–151.

3 Wang S, Li W, Wang Y, *et al*. An improved difference of gaussian filter in face recognition. Journal of Multimedia, 2012, 7(6): 429–433.

4 Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. 2005. 886–893.

5 Lowe DG. Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision. Kerkyra, Greece. 1999. 1150–1157.

6 Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971–987. [doi: 10.1109/TPAMI.2002.1017623]

7 Chen PH, Lin CJ, Schölkopf B. A tutorial on ν-support vector machines. Applied Stochastic Models in Business and Industry, 2005, 21(2): 111–136. [doi: 10.1002/asmb.v21:2]

8 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

9 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. arXiv:1311.2901, 2014.

10 Donahue J, Jia Y, Vinyals O, *et al*. DeCAF: A deep convolutional activation feature for generic visual recognition. International Conference on Machine Learning. Beijing, China. 2014. 647–655

11 Smirnov EA, Timoshenko DM, Andrianov SN. Comparison of regularization methods for imagenet classification with deep convolutional neural networks. Aasri Procedia, 2014, 6: 89–94. [doi: 10.1016/j.aasri.2014.05.013]

12 Zhou BL, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 2921–2929.

13 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. arXiv:1610.02391, 2017.

14 Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 248–255.

15 Zhou BL, Lapedriza A, Khosla A, *et al*. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452–1464. [doi: 10.1109/TPAMI.2017.2723009]

16 Herranz L, Jiang SQ, Li XY. Scene Recognition with CNNs: objects, scales and dataset bias. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. 2016. 571–579.