

引用格式:董柏岩,李熹桥,金鑫.图像美学质量评价模型的可解释性分析[J].中国传媒大学学报(自然科学版),2022,29(03):36-42.

文章编号:1673-4793(2022)03-0036-07

图像美学质量评价模型的可解释性分析

董柏岩,李熹桥,金鑫*

(北京电子科技学院,北京 100070)

摘要:在图像美学质量评价的发展过程中,前期有大量的科研工作用于人工设计美学特征,当前深度卷积网络的应用虽然减少了人工设计美学特征的流程,但使模型变得不可解释,这影响了人类对美学的认知。本文利用现有的可解释算法,对美学单一数值评价模型做出一定的解释。通过相关性归因方法,将原始输入图像对分类结果影响权重大的区域进行可视化,对证明模型的可靠性、发现新的美学规则具有十分重要的意义。

关键词:深度学习;图像美学;可解释性

中图分类号:TP391

文献标识码:A

Interpretability analysis of image aesthetic quality evaluation model

DONG Boyan, LI Xiqiao, JIN Xin*

(Beijing Electronic Science and Technology Institute, Beijing 100070, China)

Abstract: In the development process of image aesthetics quality evaluation, a large amount of scientific research work was used to artificially design aesthetic features in the early stage. Although the current deep convolutional network reduces the manual design process, the model becomes unexplainable, which affects the aesthetics of human beings' cognition. Therefore, in this article, by using the existing interpretable algorithm, a certain explanation is made on the aesthetic single numerical evaluation model. This article uses the gradient attribution method to visualize the areas where the original input image has a significant influence on the classification results, which is of great significance for proving the reliability of the model and discovering aesthetic rules.

Key words: deep learning; image aesthetics; interpretability

1 引言

图像美学质量评价是利用计算机模拟人类对美的感知与认知,自动评价图像的“美感”,分析图像在构图、颜色、光影、景深、虚实等美学因素影响下所形成的美感刺激^[1]。最初的研究基于人工设计的美学特征,通过设计的视觉特征,拟合人类对图像美学质量的评价,进行图像的美感分类与评分^[2-6]。2014年以

来,图像美学质量评价的研究工作进入了深度学习时代,研究人员改造了多种用于图像识别的卷积神经网络,并且将图像风格、图像内容等额外信息显式或隐式地建模于改造后的卷积神经网络中,这种方式构建的图像美学质量评价模型取得了更优秀的效果^[7-11]。

神经网络可解释是人工智能3.0的新要求,目标在于使得人工智能算法更加透明可靠,特别是在医疗、金融和军事领域。深度学习模型在训练过程中,

基金项目:北京高校“高精尖”学科建设项目(20210051Z0401)

作者简介(*为通讯作者):董柏岩(1998-),男,硕士研究生,主要从事人工智能技术的研究;李熹桥(1997-),男,硕士研究生,主要从事人工智能研究技术的研究;金鑫(1983-),男,副教授,博士,主要从事计算美学、计算机视觉、人工智能与安全等技术的研究。Email:jinxin@besti.edu.cn。

使用了大量的数据优化模型参数,但是神经网络中间神经元个数、网络的深度、全连接层的设置等神经网络参数会如何影响结果还需要依赖专业研究人员的工作经验。目前基于深度卷积神经网络图像设计的美学评价模型可解释性不强,这不利于图像美学评估的进一步研究。

本文使用深度学习解释性研究中事后解释性的几种方法对图像美学单一评价模型进行一定的解释,包括特征图可视化、类激活图可视化、层级相关性传播方法。通过特征图可视化理解了浅层和深度网络模型的学习重点。对模型类激活图进行可视化,证明了经过注意力模型,网络模型更加关注图像主体。通过相关性归因算法,发现了当前模型对图像中光影、轮廓等高级语义信息敏感度较强。通过对图像美学评价模型进行解释,既可以获得美学指导,也可以根据网络的特性对模型进行修改。

2 可解释性方法简介

在许多计算机视觉任务中,卷积网络的性能已经远远超过传统模式识别算法。但相比传统算法,深度卷积网络可解释性低。在一些关键领域,如医疗、航空和军事领域,许多实际应用都要路经模型背后的决策过程和决策依据,才能避免一些严重威胁。如果飞机巡航程序缺乏合理的解释性,可能会在一些环境突变的情况下给旅客造成生命危险。在日常生活中,手机面容解锁已经成为公众接受的方式,但是一些研究表明,构造对抗样本即可将面容识别模型的结果更改,这其实也是因为面容识别模型的可解释性低。

为了使人工智能算法更加透明和可靠,更安全地将模型落地,近年来,许多科研人员针对模型可解释性提出了不同的方案。这些方案可以分为事前解释和事后解释两种。事前解释一般是通过在算法设计时加入可解释模块进行的。深度卷积神经网络由于隐含层较多,透明性低。通过在网络结构引入注意力机制是一种行之有效的可解释性方法。注意力机制是对人脑认知方式的模拟,由于大脑信息处理水平的局限性,人类会对输入信息进行筛选,处理信息中重要部分,忽略不必要信息。神经网络的注意力机制是建立对输入的注意矩阵,重要部分权重大,通过反向传播,注意矩阵也会不断调整,最终体现了对结果影响最大的区域。比如在自然语言处理领域,Bahdanau^[12]等人将注意力机制引入到基于编码器-解码器架

构的机器翻译中,有效地提高了“英语-法语”翻译的性能。在编码阶段,机器翻译模型采用双向循环神经网络将原语言编码到向量空间中;在解码阶段,注意力机制为解码器的隐藏状态分配不同的权重,从而允许解码器在生成法语翻译的每个步骤选择性地处理输入句子的不同部分。最后通过可视化注意力权重,用户可以清楚地理解一种语言中的单词是如何依赖另一种语言中的单词进行正确翻译的。

事后解释主要是在模型训练结束之后,通过可视化、参数分析对比实验解释模型。分为两类,一类是全局性解释,另一类是局部性解释。全局性解释通过可视化解释每个神经元的状况来理解神经网络的有效性;局部性解释主要是理解图片分类的原因,只关注输入图像的部分区域。

针对神经网络的常见全局性解释方法有模型蒸馏、激活最大化等。深度卷积网络的复杂度高,从全局进行解释的难度高。模型蒸馏是将复杂的大模型学习到的知识迁移到简单的小模型,其目标是在保证模型泛化能力的前提下将复杂模型进行浓缩。小模型可以采用可解释性强的模型,比如决策树、浅层神经网络。模型蒸馏主要的难题是表达能力弱的小模型难以将从大量数据得到复杂特征进行表达,因而导致小模型性能较低。激活最大化是将卷积神经网络中间的特征图进行可视化,从图像的视觉语义信息直观地理解神经网络的内部逻辑。常见的局部解释算法主要有 LIME^[13]、Scaliency Maps^[14]、LRP^[15]、DeepLIFT^[16]几种,其中 LIME 是一种和模型无关的局部可解析性算法。主要思路是基于想要解释的部分模型及样本来构建局部的简单模型。该方法适用于特征易于解释的情况。具体做法是对样本进行可解释的扰动,得到若干个扰动样本,然后再将这些样本还原到特征空间中,建立局部模型。比如在动物分类任务中,只保留动物的头部或者尾巴,生成新样本。然后对只保留头部的图片作出预测,生成真实值。最后将新样本和所对应真实值进行简单建模,观察样本对于结果的影响。

3 美学单一评价模型简介

图像美学质量评价从美学特征提取方式来看可分为基于传统视觉算法构建美学规则和基于深度学习构建美学模型两个阶段。前期的图像美学质量评估主要是基于人类对图像特征和摄影规则的审美感知来人工

设计特征,包括三分法则^[17],图片的景深^[18]和平衡性^[19]等。自深度学习在计算机视觉其他领域的任务表现出色之后,国内外研究者开始使用深度神经网络来提取美学特征^[20-22],虽然可解释性比人工设计要低,但是基于深度学习的图像质量评价的效果得到了大大的改进。

目前图像美学质量评价的研究主要集中在以下两类,一类为美学分类、美学单一数值评价、美感分布预测等数值型评价研究;一类为与美学语言评论生成相关的自然语言处理研究。其中,单一数值评价是给出原始输入图像的美学评分或者属性评分。本文所介绍的图像

美学单一评价模型基于深度学习,包括了注意力机制以及美学评分分类化两个部分。注意力机制是对人类视觉机制的有效模拟,在自然语言处理中,句子中的“美好”和“乐观”等词语比“生活”这类词更有意义地决定了评论的情感。在神经网络中增加注意力模块可以有效提高模型的表达能力和模型可解释性。引入美学评分分类化则是为了美学分类指导美学回归任务。主要的依据是美学分类是弱分类,类别之间不像物品识别一样具有严格的界限。因此采取先分类再回归的方法,可以提高美学单一数值评价的性能。

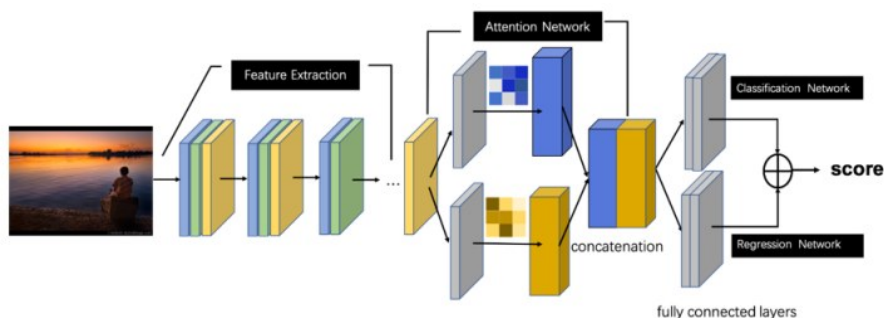


图1 图像单一数值评价网络模型示意图

如图1所示,神经网络模型主要分为三部分,分别为:特征提取部分、注意力部分以及分类回归部分。特征提取部分采用效率高的EfficientNet^[23],注意力部分采用位置注意力和通道注意力相结合的方式,分类回归部分的设计为先分类再回归。

EfficientNet是一种通过网格搜索深度、宽度和分辨率寻找到最佳比例的网络,其效率高、速度快。对于图像美学而言,更多参数的卷积神经网络将有助于多主体的图片特征的提取。模型使用双路的并行注意力机制,包括位置注意力和通道注意力。原因是图像美学质量评估中,图像中各个物体的位置及相对关系和图像整体的颜色、光照、深度都会对图像美学质量结果有较大的影响。因此使用建立图像的特征关系的位置注意力和通过将不同通道的依赖性进行学习来增强特征图的表达能力。在特征提取部分的特征图通过注意力模块之后,通过级联的方式将特征聚合,综合了多个位置、多个通道的多尺度信息。分类回归部分是在聚合特征之后使用全局池化层(Global Average Pooling, GAP),主要原因是使用全连接层会在迭代若干次之后使得模型出现过拟合的现象,而使用GAP层可以实现有效的抗过拟合效果,准确率变得稳定。位置注意力网络模型如图2所示。

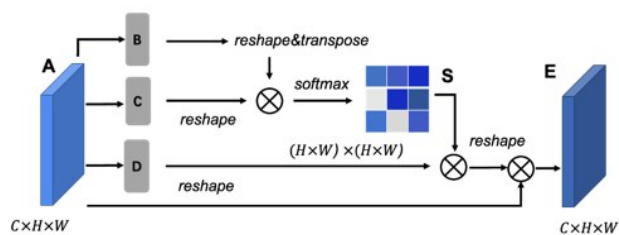


图2 位置注意力示意图

4 神经网络的特征图可视化

卷积网络在训练中学习了图像中的各种空间结构,将特征图进行可视化有助于更好地理解深度网络。在本节中,将直接可视化网络每层的特征图,观察特征图随着网络深度的变化,并且通过转置卷积在输入图像中得到映射。

转置卷积操作能够使特征图在原始输入找到对应位置,主要是因为神经网络在前向传播时,每个卷积的输出都依赖输入的区域。如图3所示,Layer1绿色区域以外的值不会影响Layer2绿色区域的值。

卷积的前向传播就是转置卷积层的反向传播过程,两者正好是相反的。卷积层可以理解为一个稀疏矩阵,卷积核以外的都为0。卷积核内部为需要学习的参数。

假设输入图像为A,卷积层为X,那么卷积后的结果为A⊗X=B。

反卷积操作的主要难点是,一般来说,为了提高模型的泛化能力,卷积网络会在卷积层之后使用池化操作来降低卷积层输出的特征向量。最大池化方法是最常用的池化操作之一,是一种不可逆的操作。因此在反卷积之后需要近似的进行反池化,也就是将特征图最大激活值的位置进行记录,其他区域赋值为0。

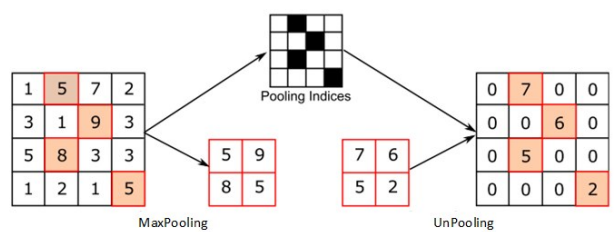


图4 反池化示意图

图5为特征图可视化的流程图,输入的图像为一幅沙漠风景图,图片内容显示正值中午,有几个人骑着骆驼在沙丘的顶峰行走。通过可视化特征层,并将特征层

映射到原始输入图像中,可以发现网络浅层部分确实将沙漠、天空、甚至阴影部分的纹理学习到了;而深层部分关注的是沙漠的流动性,并学习到了背景和前景的区别。

图6为有代表性的特征图。可以得出以下几条规律:

- (1) 浅层网络主要提取纹理、颜色、细节特征。
- (2) 深层网络主要提取轮廓、形状特征。

根据可视化的结果与对感受野的理解,得出结论:出现这种情况的原因是浅层网络感受野较小,对应的像素区域也较小,能够利用更多的细粒度特征信息捕获更多细节。而深层网络,由于下采样或卷积次数逐步增加,感受野逐渐增加,感受野之间的重叠区域也不断增加,此时的像素点代表的信息是一个区域的信息,获取的是这块区域或相邻区域之间的特征信息,对应语义信息,细粒度相对不够。特征图可视化不能完全解释神经网络,但通过这种方法可以发现卷积神经网络学习的特征是呈现分层特征,浅层网络抽取图像的具体特征,而深层网络呈现图像抽象特征,这个过程与人类认知一致。这对未来设计美学深度模型提供了指导,比如针对图像色彩评价可以利用浅层特征,有效防止过拟合。

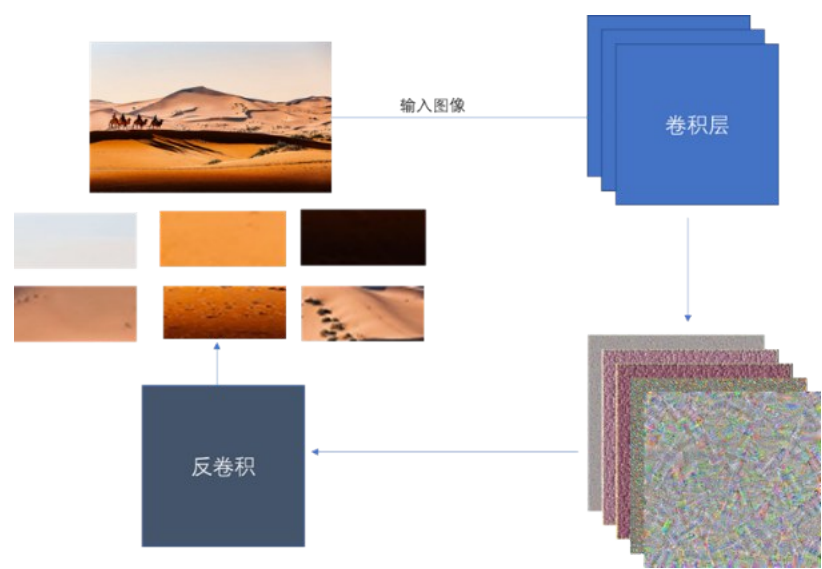


图5 特征图可视化流程

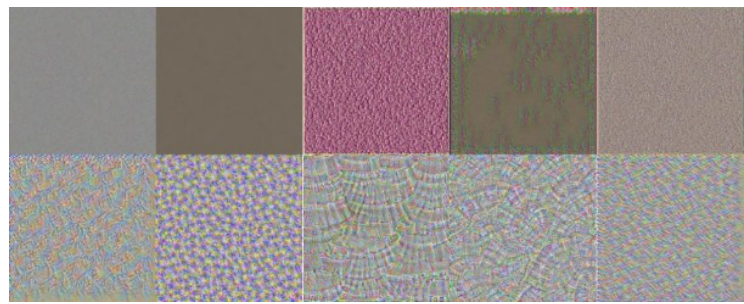


图6 代表性特征图

5 类激活图的可视化

类激活图可视化,是一种神经网络解释方法,通过对输入图像生成类激活的热力图来表示每个位置对该类别的重要程度,有助于通过可视化把握输入图片的不同区域对预测结果的影响程度。

类激活图利用特征图权重叠加的原理进行可视化。具体而言,就是将神经网络结构中的全连接层更换为全局平均池化加全连接层。全局平均池化层的工作原理是将池化层的滑动窗口设置为与特征图的尺寸一致,这种做法可以理解为将特征图每一通道压缩为一个值。在神经网络中使用全局平均池化层可以使得模型参数大大减少,运行速度更快,有效防止过拟合。

类激活的主要公式为:

$$L_{CAM}^c = \sum_{i=1}^n w_i^c A^i \quad (1)$$

重新推导此过程: A 表示网络最后的卷积层输出,大小为 $w \times h \times n$,其中 w 为最后小全连接层的权重, C 为类别数。经过一个全局池化层得到 $1 \times n$,然后经过一个权重为 $n \times c$ 的小全连接层,得到大小为 $1 \times c$ 的分类向量。最后高亮图片中对分类最重要作用的区域,得到图像的类激活图。

在本节中使用类激活图可视化方法评测美学单一评价模型中注意力机制的有效性。图7中第一列为原始输入图像,第二列为输入图像只通过特征提取层的结果,第三列为输入图像通过注意力机制后的特征图。颜色越靠近红色,说明影响越大;颜色越蓝则说明影响越小。可以发现,在经过位置注意力和通道注意力之后,神经网络可以定位到主体上,且主体对结果的影响增大。合理利用注意力机制可以有效提升神经网络性能。

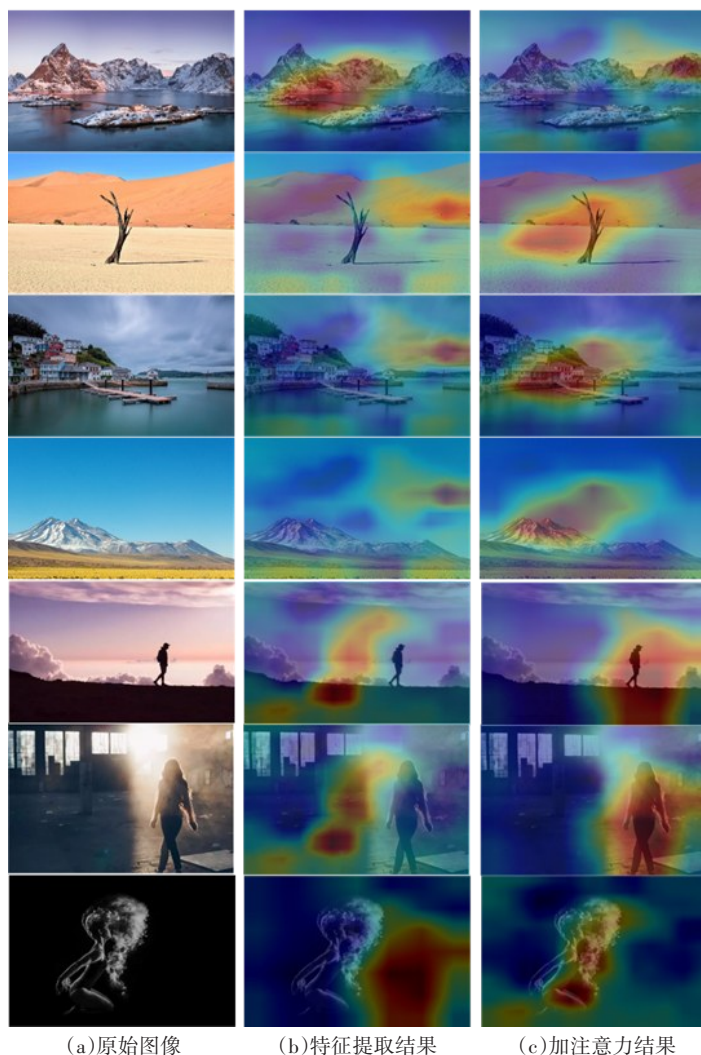


图7 类激活图结果

6 相关性归因

基于梯度的方法对卷积神经网络的结果进行归因,是可解释性神经网络的主要方法。一个成熟的归因方法要满足敏感性和实现不变性。敏感性是一个归因方法对于所有的输入和基准输入都应该满足敏感性,即对不同的输入特征,产生不同的预测结果时,不同的特征所对应的归因(属性)非0。实现不变性是指如果两个完全不同的方式实现的网络对于所有输入、输出都相等,则两个网络在功能上是等效的。

层级相关性传播方法是一种试图将分类器结果在原始输入图像上找到对应关系,得到图像中每个像素的决策权重的方法。层级归因化是一种贡献传播

方法,该方法首先利用加性模型计算高层特征对模型预测结果的贡献,然后通过反向传播将高层特征的贡献逐层传递到模型的输入,以确定每一层的每一个神经元节点对其下一层神经元节点的相对贡献。该方法将模型高层特征分解为多个多维向量,且每个多维向量对应一个相关性分值,然后将相关性分值递归地传播到低层。基于梯度反向传播的解释方法虽然充分利用了模型的结构特性,但无法实际解决预测函数梯度小导致激活函数失真的问题,因此这类解释方法依然存在许多人类无法理解的噪音。而基于重要性或相关性的反向传播方法不要求激活是可微的,可以有效解决该问题。

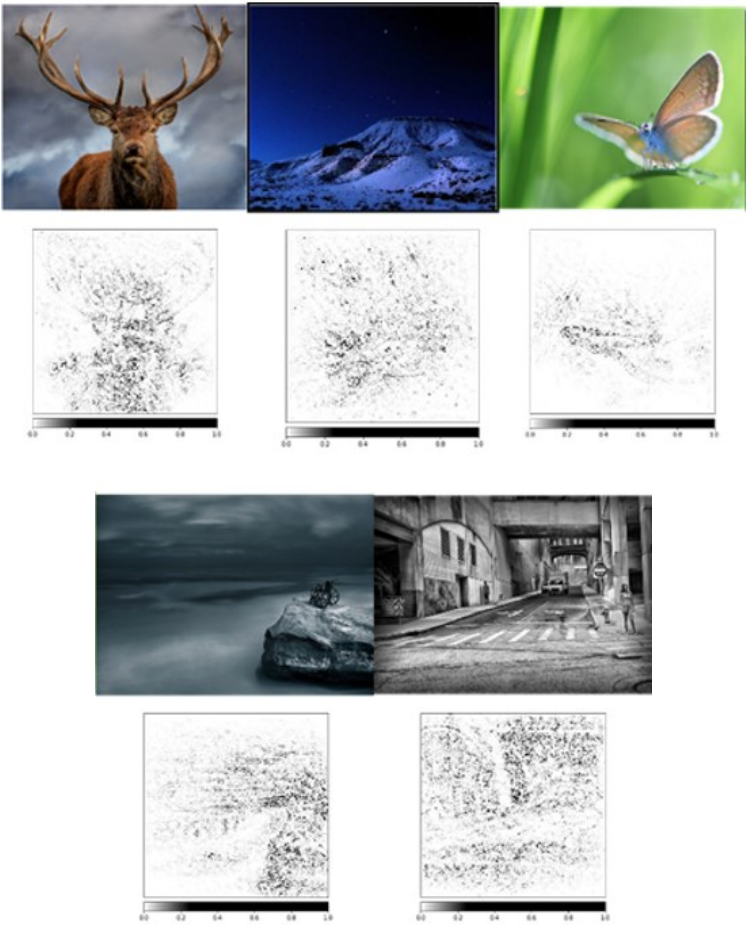


图8 层级相关性结果

本节运用层级相关性传播方法计算标签与输入图像的关联,并将结果可视化。如图8显示,原始输入图像中光影、轮廓等信息对预测结果影响权重大。第一幅图画面前景部分为动物,背景为乌云,通过层级相关性可以发现这类肖像图前景主体对结果的影响比背景大。

从第二幅图像和第三幅图像的结果可以看出光影变化的位置比图像中同色调的部分对结果的影响权重大。第4幅图是一把轮椅在悬崖上,轮椅的位置在图像的三分位置,画面中其余部分表现了一种静谧阴森的感觉。从可视化结果来看,模型可以准确的识别画面的主体。

7 总结

本文运用当前深度学习领域前沿的可解性方法对图像美学单一数值模型进行解释。主要使用了基于梯度的类激活图可视化、上卷积网络提取特征图、层级归因算法三种方法,并对三种方法展现的结果进行解释。进行可解释性工作一方面证明了模型性能变好的依据,比如注意力机制使得卷积网络更加关注图像中主体,而不是关注图像中无用信息。一方面在解释模型之后可根据输出结果的特性对模型进行改进,同时在模型解释的过程中可以提供对美学的指导。

参考文献(References):

- [1] 金鑫,周彬,邹冬青,等. 图像美学质量评价技术发展趋势[J]. 科技导报, 2018, 36(9): 36-45; DOI: 10.3981/j.issn.1000-7857.2018.09.005.
- [2] Tong H, Li M, Zhang H, et al. Classification of digital photos taken by photographers or home users[C]. Proceedings of the 5th Pacific Rim Conference on Advances in Multimedia Information Processing. Heidelberg: Springer-Verlag, 2004: 198-205.
- [3] Luo Y, Tang X. Photo and video quality evaluation: Focusing on the subject[C]. Proceedings of the 10th European Conference on Computer Vision (ECCV). Heidelberg: Springer Verlag, 2008: 386-399.
- [4] Wong L K, Low K L. Saliency-enhanced image aesthetics class prediction[C]. Proceedings of the 16th IEEE International Conference on Image Processing. Piscataway NJ: IEEE, 2009: 997-1000.
- [5] Li C, Chen T. Aesthetic visual quality assessment of paintings [J]. IEEE Journal of Selected Topics in Signal Processing, 2009, 3(2): 236-252.
- [6] Nishiyama M, Okabe T, Sato I, et al. Aesthetic quality classification of photographs based on color harmony[C]. Proceedings of the 24th IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway NJ: IEEE, 2011: 33-40.
- [7] Lu X, Lin Z, Jin H, et al. RAPID: Rating pictorial aesthetics using deep learning[C]. Proceedings of the ACM International Conference on Multimedia. New York: ACM, 2014: 457-466.
- [8] Lu X, Lin Z, Shen X, et al. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation[C]. Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway NJ: IEEE, 2015: 990-998.
- [9] Wang W, Zhao M, Wang L, et al. A multi-scene deep learning model for image aesthetic evaluation[J]. Signal Processing Image Communication, 2016, 47(C): 511-518.
- [10] Kong S, Shen X, Lin Z, et al. Photo aesthetics ranking network with attributes and content adaptation[C]. Proceedings of 14th European Conference on Computer Vision. Heidelberg: Springer-Verlag, 2016: 662-679.
- [11] Ma S, Liu J, Chen C W. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway NJ: IEEE, 2017: 722-731.
- [12] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint, arXiv: 1409.0473, 2014.
- [13] Ribeiro M T, Singh S, Guestrin C. "Why should i trust you? Explaining the predictions of any classifier[C]. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 1135-1144.
- [14] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv preprint, arXiv:1312.6034, 2013.
- [15] 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.
- [16] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences [C]. International Conference on Machine Learning. PMLR, 2017: 3145-3153.
- [17] Obrador P, Anguera X, de Oliveira R, et al. The role of tags and image aesthetics in social image search[C]. Proceedings of the first SIGMM workshop on Social media. 2009: 65-72.
- [18] Datta, Ritendra, Jia Li, and James Ze Wang. "Studying aesthetics in photographic images using a computational approach." european conference on computer vision (2006): 288-301
- [19] Ke, Yan, Xiaoou Tang, and Feng Jing. "The Design of High-Level Features for Photo Quality Assessment." computer vision and pattern recognition (2006): 419-426.5
- [20] Jin B, Segovia M V O, Süssstrunk S. 2016. Image aesthetic predictors based on weighted CNNs[C]. Proceedings of 2016 IEEE International Conference on Image Processing. Piscataway NJ: IEEE, 2016: 2291-2295.
- [21] Jin X, Wu L, Li X D, et al. ILGNet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation[J]. arXiv. org, 2018, arXiv:1610.02256v3.
- [22] Wang Z, Liu D, Chang S, et al. Image aesthetics assessment using deep chatterjee's machine[C]. Proceedings of 2017 International Joint Conference on Neural Networks. Piscataway NJ: IEEE, 2017: 941-948.
- [23] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

编辑:龙学锋