基于深度学习的少数民族绘画情感分析方法

张浩, 徐丹*

云南大学信息学院, 昆明 650504

* 通信作者. E-mail: danxu@ynu.edu.cn

收稿日期: 2018-09-10; 接受日期: 2018-10-29; 网络出版日期: 2019-02-18

国家自然科学基金 (批准号: 61163019, 61540062)、云南省应用基础研究重点项目 (批准号: 2014FA021) 和云南省教育厅科学研究基金产业化培育项目 (批准号: 2016CYH03) 资助

摘要 图像不仅可以传达丰富的语义信息进而直接表达人的情感, 也可以在情感层面影响他人, 艺术作品更是如此. 人们对于不同的视觉刺激有着不同的情绪反应, 所以理解图片所包含的情感是非常重要的. 本文通过微调现有的卷积神经网络模型, 以少数民族绘画图像数据为研究对象, 分析了艺术绘画的色相、亮度、饱和度, 以及对比度对神经网络模型情感分类的影响, 并在公共数据集 Twitter image dataset 上取得了超过目前先进水平 3.4% 的精度提升. 最后, 本文提出了一种相关任务的预训练策略, 对少数民族绘画的情感分类任务有明显的提升, 并通过可视化结构对实验结果进行了分析.

关键词 少数民族绘画, 图像情感, 卷积神经网络, 微调, 可视化

1 引言

随着互联网时代的发展, 多媒体社交网络也被越来越多的人所广泛使用, 例如 Facebook、Flickr、Twitter、新浪微博等, 越来越多的人习惯于记录下他们的日常生活并利用社交媒体分享出来, 其手段包括文本、音频、图片, 以及视频等. 公众对视觉内容情感表达的强烈需求, 使得对图像、视频情感层面的分析变得越来越迫切. Minsky 教授的《The Society of Mind》最早提出拥有情感的智能机器, 但情感计算的概念是由 Massachusetts Institute of Technology 媒体实验室的 Picard 教授于 1995 年提出. 1997 年专著《Affective Computing》正式出版^[1], 给出情感计算的定义, 即与情感相关、来源于情感或者能够对情感施加影响的计算, 情感计算的思想是要使得计算机能够观察、理解和表达各类情感特征等, 最终像人一样能随意的表达和交流. 对于多媒体数据的情感分析, 一方面, 基于文本的情感分析^[2]和基于音频的情感分析^[3]技术成熟, 而基于图像的情感分析^[4,5]以及基于视频的情感分析^[6]还未引起大多数研究者的注意; 另一方面, 图像情感计算的发展离不开心理学、艺术学、计算机视觉、模式识别、人工智能等领域的共同支持, 交叉学科带来的挑战也使得图像情感分析极具挑战; 再者, 目前还

引用格式: 张浩, 徐丹. 基于深度学习的少数民族绘画情感分析方法. 中国科学: 信息科学, 2019, 49: 204–215, doi: 10.1360/N112018-00249
Zhang H, Xu D. Ethnic painting analysis based on deep learning (in Chinese). Sci Sin Inform, 2019, 49: 204–215, doi: 10.1360/N112018-00249

缺乏对绘画艺术作品进行比较系统的情感语义研究. 而作为人类的精神追求, 艺术常常以丰富的情感、风格迥异的绘画技法, 影响观赏者对图像的情感体验. 因此, 图像情感计算在目前的研究中越来越显现其至关重要的地位. 本文结合云南省丰富的少数民族绘画资源, 分析艺术特色的情感问题.

云南地处我国西南边陲, 东邻广西贵州, 北接四川, 西依西藏、缅甸, 南部更是与老挝、越南毗邻, 国境线长达 4060 km, 是我国通往南亚、东南亚的窗口和门户. 自然风光绚丽, 大家耳熟能详的丽江古城、苍山洱海、西双版纳热带雨林等均在云南. 云南以其独特的自然条件和地理环境, 成为一个多民族边疆省份、众多民族生息繁衍之地, 26 个民族齐聚于此, 其悠久的历史背景和灿烂的文化底蕴, 孕育了民族特色鲜明、熠熠生辉的非物质文化遗产, 素有“彩云之南”的美称. 而少数民族绘画艺术方面, 有着各自鲜明的民族特色, 在那些发展比较缓慢的民族中, 其民族美术往往带有“历史化石”的性质, 而较为先进的民族, 则是将先进文化融入其中, 形成别具一格的绘画艺术创作. 云南少数民族绘画, 深刻地反映了云南民间绘画题材所具有的民俗文化和民族特征, 较好地体现了少数民族的审美视觉. 本文正是在这种图像情感计算愈渐迫切的需求情况之下, 从少数民族艺术绘画入手, 以包含云南版画、油画、重彩画、水彩, 以及水粉画的数据集为基础, 针对少数民族绘画的艺术情感进行分析和讨论.

2 相关工作

虽然基于图像和视频的情感分析相对于基于文本和音频的情感分析还比较少, 但也引起了一些研究者的注意. 文献 [7] 以颜色为主要特征, 提出了一种基于模糊相似性的彩色图像情感分类方法. 文献 [8] 通过形状特征研究情感的可计算性. 文献 [9] 提取包括了图像的细节程度、动态性、低景深, 以及三分法的构图作为其情感分类的中层特征. 文献 [10] 使用各个兴趣点周围的情感直方图特征以及情感包特征对图像进行情感分类. 还有一些研究者则是结合艺术元素特征进行图像情感的分析, 如文献 [11] 就构造了对称、强调、和谐、层次、运动、节奏和比例 6 个基于艺术元素的特征对图像情感分类. 而随着计算机硬件计算能力的迅速发展, 深度学习在各个领域内被广泛使用, 文献 [12] 也提及其特征具有极强的表示能力, 其中一些研究者们则是将深度学习用于图像情感计算中, 避免了人为抽取特征的方式, 取得了不错的进展. 文献 [13, 14] 进行了图像情感的二值分类. 文献 [15~18] 则是将图像情感分为娱乐、生气、敬畏、满意、厌恶、刺激、害怕和悲伤 8 类. 但总的看来, 还缺乏对绘画艺术作品进行比较系统的情感语义研究.

近年来, 基于卷积神经网络的深度学习方法在许多领域都取得了重大的成功, 而且被广泛应用于各个领域, 在计算机视觉领域内, 相比于手工提取特征的效果也是立竿见影, 如分类问题^[19~21]、物体检测^[22~24]、图像语义分割^[25]等. 其特征层通过不断地训练数据进行学习, 避免了显式的特征抽取, 隐式地从训练数据中寻找规律进行学习, 这就排除了手动提取的特征的局限性, 特别是包含多维信息的图像数据可直接输入网络, 也可以有效避免学习过程中特征提取和分类过程中数据重建的复杂度, 而且同一个特征层的同一映射面上的神经元权值相同, 使得网络得以并行学习, 减少计算量的同时缩短了学习周期, 大大提升了学习效率. 对于图像情感分类任务这个复杂的问题, 卷积神经网络与手工提取特征方法不同的是, 后者往往更侧重于如何选取从图像中抽取的特征来表示其要反映的情感, 而采用卷积神经网络之后算法本身则更侧重于模型的建立, 这就排除了手动提取的特征的局限性, 并且使得底层特征和高层语义之间的映射成为可能.

本文结合云南省丰富的民族绘画资源, 采用卷积神经网络作为实验模型, 针对云南省少数民族绘画的艺术情感进行分析和讨论. 与文献 [13] 的情感分类一致, 本文将艺术图片分为积极的情绪和消极的情绪两类, 采用微调这种迁移学习的策略, 解决数据集不是足够大时可能出现的过拟合的问题.

为了使得模型具有更好的泛化能力, 往往采用适当的方式进行数据集扩充以增强模型的学习能力^[26], 本文针对民族艺术绘画和普通图像的情感分类任务, 分别分析了几种常用数据集扩充方式对于图像情感问题中网络性能的提升, 包括裁剪、翻转, 以及图像的色相、亮度、饱和度和对比度, 然后选取合理的组合进行过采样并运用到本文实验, 避免了盲目地凭借直觉进行随意地修改数据. 为了验证本文模型的性能的有效性, 我们在公共数据集 Twitter image dataset 上与目前最先进水平的方法^[13,14]进行了比较, 并取得了 3.4% 的性能提升.

最后, 为了研究卷积神经网络对于不同计算机视觉任务的影响, 即在不同计算机视觉任务中的普适性, 我们经过大量实验分析, 提出了一种相关任务的预训练策略以提升模型性能, 提升模型对民族绘画的情感分类问题的预测, 并使用可视化的分析体现了模型学习情况的可解释性. 本文模型开源代码见网站¹⁾.

3 方法和实验

3.1 微调 MXNet 模型用于民族绘画情感分类

作为挖掘数据价值的一大助力, 卷积神经网络模型的训练离不开数据的依托, 而微调这种策略不仅仅能够解决没有大数据集的问题, 而且能够使得神经网络的训练时间大大下降, 很大程度地解决了数据集不大时过拟合的问题. 微调不仅使得小数据集采用深度学习成为可能, 而且往往能够取得不错的效果^[27~29]. 一方面, 不同于普通图像, 民族绘画作为艺术绘画的一个分支, 本身数量有限; 另一方面, 对图像的标注更需要一定的要求, 标注人员需要有一定的艺术素养才能更为精准地理解和诠释图像的情感, 使得收集数据的过程变得异常艰难. 因此, 在数据量有限的情况下, 本文采用预训练好的卷积神经网络 VGG16^[21] 作为本文的实验模型, 对少数民族艺术绘画进行情感的分类.

实验采用深度学习框架 MXNet, 使用 ILSVRC2012 上预训练好的模型 VGG16 作为基本结构, 去除原网络最后一个包含 1000 个神经元的全连接层, 添加包含 2 个神经元的全连接层作为情感预测的结果, 以此适应图像情感的二值分类任务.

3.2 过采样提高模型泛化能力

为了充分利用已经收集好的数据训练卷积神经网络模型, 也为了使模型具有更强的鲁棒性, 我们对原图像进行适当的改变以扩充数据集训练模型. 文献^[26]提出了一种扩充数据集训练网络的方法以加强网络的学习后的泛化能力, 如翻转或者裁剪等. 为了验证这种方法在民族绘画艺术图像任务上对于模型性能的影响, 我们采用几种常见的扩充数据集方式对卷积神经网络进行训练, 包括图像的裁剪、翻转, 以及图像的色相、亮度、饱和度和对比度. 一般来说, 裁剪、翻转使得感兴趣的物体出现在不同的位置中, 从而使得模型减小对物体出现位置的依赖性, 调整色彩等因素来降低模型对色彩的敏感度. 裁剪时采取随机裁剪的方式但保留原图像至少 70% 的区域, 同时高和宽的比例控制在 [0.75, 1.25] 之间, 翻转则是让每张图像以 0.5 的概率左右翻转, 不采取上下翻转是因为我们在现实生活一般不会对倒立的图像产生兴趣, 更不会试图去理解倒立图像的情感问题. 对于绘画图像的色相、亮度、饱和度, 以及对比度 4 个维度, 每个维度做随机 -50%~50% 的增量, 如图 1 所示. 按照以上几种方式分别对民族艺术绘画数据集进行扩充, 重新对微调 MXNet 模型进行训练, 最后选取对模型性能有提升的几种方式进行组合, 分析有效的数据扩充方式对于模型性能的具体提升情况.

1) <https://github.com/haoyev5/Ethnic-Painting-Analysis>.

图1 (网络版彩图) 部分过采样方式示例

Figure 1 (Color online) Examples of oversampling mode. (a) Original image; (b) random cutting; (c) image brightness changed; (d) image color changed

最后, 将本文模型在公开的数据集 Twitter image dataset 上做了实验和分析, 进而验证本文模型的有效性. 同时, 为了分析不同数据扩充方式在艺术绘画情感分类和普通图像情感分类任务之间的差异, 将以上 6 种不同的数据扩充方式以及有效的组合方式在普通图像情感分类任务上也进行了类似的实验和分析.

3.3 相关任务的预训练策略

虽然微调作为一种非常有效的迁移学习方法, 不用重新开始训练而且一般情况下可以快速收敛, 但是通常情况下, 微调模型初始化权重和偏置时, 除去因不同的视觉任务而更换的最后一层, 需要保留其他层的参数, 因此可能会导致卷积神经网络将学习到的部分与此任务不相关的经验带入到其中, 最终干扰模型的性能. 而本文涉及到的少数民族艺术绘画的情感识别问题, 相对于普通照片而言更具有挑战. 首先, 艺术风格的图片往往更为抽象, 观者往往需要有一定的背景才能准确把控, 导致数据标签的模糊度较高, 其次, 民族风格绘画相对于普通图像, 其蕴含的感情更为复杂, 细节上的描述也更为具体, 从而对卷积神经网络提出更高的要求, 这就为艺术绘画的情感识别任务带来了更大的挑战. 我们试图模仿人类的学习行为, 学习知识的难度应当由简入繁、由浅入深. 因此, 本文提出一种新的相关任务的预训练策略 (pre-training strategies for related tasks), 让模型先适应相对而言更为容易但是相关的任务, 再将其用于要处理的相对更具有挑战性的识别问题.

我们将训练好的 MXNet 在一个相对而言更容易学习的图像情感分类问题上进行预训练, 然后再将此模型用于民族绘画的情感分类问题上, 以此模拟人类由浅入深的学习方式, 避免过多的受已有学习经验的干扰. 因为 ILSVRC2012 是一个庞大的数据集, 在此基础上训练好的网络势必会保留大量的与之有关的特征, 这会对新的计算机视觉任务造成影响, 比如 ILSVRC2012 中蜥蜴这个类别, 有非常漂亮五彩斑斓的蜥蜴, 也有浑身坑坑洼洼令人呕吐的蜥蜴, 对于图像情感识别任务和 ILSVRC2012 上的分类问题, 显然有不同的答案.

图 2 表示了相关任务的预训练策略具体框架, 先对预训练好的 VGG16 进行微调, 将网络最后一层替换为只有两个神经元的全连接层, 让该模型在 Twitter image dataset 数据集上进行训练, 学习一些对图像情感分类问题更为有用的特征, 同时减少已有学习经验对分类结果的干扰, 然后将此模型用于少数民族绘画的情感识别任务.

3.4 预测任务的可视化

为了较为直观地展示模型的学习情况, 我们改变微调的 MXNet 模型结构, 以便可视化模型的学习

图 2 (网络版彩图) 相关任务的预训练策略, 先将微调的 VGG16 模型在 Twitter image dataset 上进行训练, 然后将此模型在少数民族绘画数据集上训练

Figure 2 (Color online) Pre-training strategies for related tasks. The fine-tuning VGG16 model was trained on the Twitter image dataset firstly, and then, trained on the dataset of ethnic painting

习情况, 为实验结果提供更加直观的可解释性. 文献 [25] 中提出用卷积层代替全连接层的思路, 不仅仅提高了计算效率, 同时也具有更为广泛的实用性, 因为在两类层中, 神经元都进行的相同的函数形式点积运算, 而两者之间的不同也仅仅是卷积层中的神经元只与输入数据中的一个局部区域连接, 并且神经元共享参数, 因此二者的转化是可能的. 我们按照这种思路改变原有的模型结构, 如图 3 所示, 将本文网络结构的前 13 个卷积层保留下来, 将 Flatten 以及后面的全连接网络替换为 3 个卷积层: Conv14 通道数 4096, 卷积核 7×7 ; Conv15 通道数 4096, 卷积核 1×1 ; Conv16 通道数 2, 卷积核 1×1 . 卷积层之间保持原来 0.5 概率的 Dropout 不变, 采用非线性函数 ReLU (rectified linear unit) 作为激活函数, 损失函数采用 Softmax 交叉熵损失. 最后, 只需要将网络的输入调整为 448×448 , 则会输出一个大小为 8×8 的预测块, 该预测块的预测值正反映了网络对于其预测的结果, 将此预测块采用邻近插值法放大至原图大小, 以此用于表示对图像区域化的情感预测.

4 实验结果和分析

本节就上部分的实验描述进行实施, 并对结果进行分析和讨论. 为便于观察, 最优结果在表格中用黑体标出.

4.1 微调模型在民族绘画数据集上的性能

在本文中, 数据集通过扫描画册中实验环境需要的绘画作品而得, 如《新中国美术 60 年》、《云南版画五十年》、《四人油画作品集》等知名画家的绘画作品集. 利用云南民族绘画情感标注系统^[30]标注图像, 该标注系统按年龄、学历、性别, 以及艺术素养的不同为每人选择不同类别属性的标记人员, 不同类别属性的标记人员对于计算最终图像情感的权重会有所不同, 如画家权重为 2, 普通标记人员则为 1, 最后根据标注情感取概率大的作为图像情感. 最终, 收集到包含云南重彩画、云南版画、云南油画, 以及水彩和水粉画等 1566 张民族艺术绘画, 其中包含 1149 幅积极情绪和 417 幅消极情绪的图像, 图片示例如图 4.

采用 MXNet 作为本文的实验框架, 微调 ILSVRC2012 上预训练好的模型 VGG16, 为适应民族绘画情感的二值分类任务, 用包含 2 个神经元的全连接层替换原有的 1000 个神经元的全连接层, 用于代表输出积极的情感和消极的情感. 除最后一层外, 全部参数均初始化为在 ILSVRC2012 上预训练好

图 3 (网络版彩图) 将网络最后的 3 个全连接层替换为 3 个卷积层: Conv14 通道数 4096, 卷积核 7×7 ; Conv15 通道数 4096, 卷积核 1×1 ; Conv16 通道数 2, 卷积核 1×1 . 最后一个卷积层的两个通道用于两类情感的预测

Figure 3 (Color online) Replace the last three full connection layers of the CNNs with three convolution layers: Conv14 contains 4096 channels, kernel size 7×7 ; Conv15 contains 4096 channels, kernel size 1×1 ; Conv16 contains 2 channels, kernel size 1×1 . The last convolutional layer is used for prediction of two kinds of emotions

图 4 (网络版彩图) 民族艺术绘画图像数据集示例, 第 1 行为积极的情感, 第 2 行为消极的情感

Figure 4 (Color online) Examples of ethnic painting image dataset, the first line behaviors positive emotion, while the second behaviors negative emotion

的参数, 替换的最后一层网络参数使用 MXNet 框架延后的初始化, 权重初始化为 $[-0.07, 0.07]$ 的均匀分布, 偏置则全部初始化为 0, 初始学习率设置为 0.001, 每 10 轮缩小 10 倍. 为充分使用数据集, 采用 5 折交叉验证将数据集分为训练集和验证集, 使用随机梯度下降的方式训练, 5 次实验后的均值作为模型的最终结果, 见表 1 第 2 行.

表 1 微调模型 VGG 在民族绘画数据集上的性能

Table 1 Performance of the fine-tuning model VGG on the ethnic painting image dataset

Model	The ethnic painting image dataset
Fine-tuning VGG16 (without oversampling)	0.701+0.020
Fine-tuning VGG16 (with oversampling)	0.723+0.013

表 2 几种过采样方式在民族绘画数据集上的性能

Table 2 Performance of several oversampling methods on the ethnic painting image dataset

Oversampling mode	The ethnic painting image dataset
Baseline	0.702+0.020
Cutting + Flipping	0.709+0.014
Brightness	0.714+0.029
Hue	0.701+0.037
Saturation	0.711+0.011
Contrast	0.693+0.041

4.2 过采样方式对模型性能的影响

本小节对比了通过对数据集进行裁剪、翻转, 改变图像的色相、亮度、饱和度, 以及对比度这几种扩充数据集的方式在图像情感分类任务中对模型性能的影响. 为提高模型的泛化能力, 采用文献 [19] 的方法对数据集进行不同方式的变化, 将改变过的数据重新反馈给网络以此来增加模型的学习能力, 按照 3.2 小节所述进行过采样, 5 次实验后的结果见表 2, 将未过采样训练模型的结果作为基准线 (baseline).

通过表 2 可以看出, 采用裁剪 + 翻转的组合、亮度, 以及饱和度 3 种方式可以有效地提升卷积神经网络模型的性能, 不仅仅精度提升, 标准差也在合适的范围内. 而适当地改变亮度对于网络性能的提升最为明显, 对于色相以及对比度两项, 会对模型的性能造成不好的效果, 这与我们的认知相符合. 改变色相即是改变颜色, 而颜色往往能够给人直观的心理暗示^[31], 例如红色代表热情、快乐、兴奋等, 而与之相邻的紫色则是代表神奇、怪异等. 至于对比度的变化反而使得模型的性能最差, 这与文献 [11] 的结论相一致, 增加或减少图像的对比度会影响模型的注意力的变化. 最终, 我们采用裁剪 + 翻转的组合, 配合增加或减小最大 50% 亮度、饱和度的变化进行过采样微调 MXNet, 结果见表 1 末行.

为了验证模型在普通图像的情感分类任务是否遵循这一原则, 我们在公开数据集 Twitter image dataset 上进行了实验. 该数据集由普通图片构成, 由亚马逊人工标注平台经 5 人标注, 有 3-agree, 4-agree, 以及 5-agree 3 个数据集, 分别包含 1269, 1115, 以及 882 张图片. 在图片标记过程中, 对于 1 张待标记的图片, 若 5 人中有 3 人给出的标签一致, 则此图片被收集到 3-agree 中, 若 5 人中有 4 人给出的标签一致, 则此图片被收集到 3-agree, 4-agree 中, 若 5 人的标记结果一致, 那么这张图片将被同时收集到 3-agree, 4-agree, 以及 5-agree 中, 其他情况下的图片直接舍弃. 为了保证数据标签的准确性, 我们仅在 5-agree 数据集上进行了实验, 以裁剪 + 翻转过采样的方式作为 baseline, 在此基础上分别对原图像亮度、色相、饱和度, 以及对比度随机增加或者减少 0~50%, 实验结果见表 3.

通过实验发现, 除了对图像进行色相的改变之外, 其他扩充数据的方式均提升了模型对于图像情感识别任务的性能, 与绘画情感识别相同的是, 亮度依旧是最为有效的方式, 而改变图像饱和度的方式在普通的图像情感识别任务上对于模型性能的提升并不明显. 这也证明了绘画图像的情感分类任务

表3 几种过采样方式在 Twitter image dataset 数据集上的性能

Table 3 Performance of several oversampling methods on the Twitter image dataset

Oversampling mode	5-agree
Baseline	0.865+0.020
Brightness	0.874+0.013
Hue	0.848+0.031
Saturation	0.867+0.017
Contrast	0.870+0.025

表4 Twitter Image 数据集上的平均分类性能和标准差

Table 4 Average classification performance and standard deviation on the Twitter image dataset

Model	3-agree	4-agree	5-agree
Baseline PCNN from [13]	0.687	0.714	0.783
Paper [14] (without oversampling)	—	—	0.839 ± 0.029
Paper [14] (with oversampling)	—	—	0.844 ± 0.026
Ours (without oversampling)	0.762+0.032	0.814+0.028	0.858+0.029
Ours (with oversampling)	0.784+0.021	0.834+0.017	0.878+0.013

和普通图像的情感分类任务具有相似性,但不像普通的图像分类问题那样能照搬相同的训练策略.为了验证这种扩充数据集方式的有效性,对原图像进行裁剪、翻转,对原图像颜色的亮度、饱和度,以及对比度进行改变,在 Twitter image dataset 上进行了测试,均采用 5 折交叉验证的方式将数据集分为训练集和测试集,并与目前最先进水平的方法 [13,14] 的结果进行对比,实验结果见表 4.

表 4 中数据可以看出,本文模型的性能要优于前两者水平,当采用有效的过采样方式时,模型在图像情感分类任务上得到进一步提升,取得了超出目前最先进水平 3.4% 的提升.纵向比较,无论是 3-agree 上的弱标签数据还是 5-agree 上的强标签数据,模型的标准差都在进一步缩小,表明模型的稳定性得到了进一步提升,在绘画艺术图像数据集上也有此规律,不仅如此,横向比较而言,数据集中数据标签越准确,标准差越低.

实验结果表明,尽管某些过采样的方式被很多卷积神经网络模型证明是非常有效的,但针对不同的计算机视觉任务,还是需要有不同的过采样策略.

4.3 相关任务的预训练策略的结果分析

为了尽可能地使预训练好的卷积神经网络模型不过多地将学习经验带入到解决新的复杂问题当中,我们先对微调的 MXNet 模型在 Twitter image dataset 上进行训练,经过前面的分析,不再采用 5 折交叉验证的方式划分数据集,而是把全部数据一并当作训练数据集,将网络最后一层替换为仅含两个神经元的全连接层.为进行实验结果的最终比较,分两组进行,第 1 组不使用过采样扩充数据集,第 2 组与 4.2 小节中的最优模型一致,采用裁剪、翻转以及最大 50% 的增量改变图像亮度、饱和度和对比度来扩充数据集进行过采样.两组实验采用相同的参数设置,批量设为 64,采用随机梯度下降训练模型,引入动量法并初始化动量参数为 0.9,初始学习率设为 0.01,每经过 10 次变为原来的 0.1,总共训练 50 轮.将得到的模型再用于民族绘画情感识别任务,使用 5 折交叉验证将少数民族绘画数据集划分为训练集和测试集,对于第 1 组,不采用数据扩充,对于第 2 组,采用裁剪、翻转以及最大 50% 的增量改

表 5 相关任务预训练策略结果

Table 5 Performance of pre-training strategies for related tasks

Model	Without oversampling	With oversampling
Fine-tuning MXNet	0.701+0.020	0.723+0.013
PS CNN	0.736+0.017	0.753+0.013

图 5 (网络版彩图) 使用 VGG16-based FCN 得出的部分结果, 第 1 行为原图, 第 2 行为生成的预测图, 最后是真实标签, 绿色代表积极的预测结果, 红色代表消极的预测结果, 颜色越浓代表网络的预测概率越大

Figure 5 (Color online) Partial results obtained by VGG16-based FCN, the first line is the original images, the generated prediction maps show in the second line, and the last is actual labels. Green represents positive predictions, while red is negative. The stronger the color, the higher prediction probability of CNNs

变图像亮度以及饱和度来扩充数据集进行过采样. 同样保持两组实验参数不变, 批量设为 64, 采用随机梯度下降训练模型, 动量法初始化动量参数为 0.9, 初始学习率设为 0.01, 每经过 15 次变为原来的 0.1, 5 折交叉验证结果的均值作为当次实验的结果. 5 次实验后的均值作为模型的最终结果, 实验结果见表 5, 4.1 小节中微调 MXNet 的实验结果作为 baseline.

实验数据表明, 这种策略对模型性能的提升要高于扩充数据集进行过采样训练模型的手段. 令人值得注意的是, 这种策略可以很好地结合过采样的方式, 共同提升卷积神经网络在具体任务上的性能, 而且从其标准差的大小可以看出, 这种策略带来的提升对于图像情感分类任务是比较稳定的.

4.4 可视化分析

我们使用训练好的 VGG16-based FCN, 将输入设置为 448×448 的图像, 经过该网络之后, 得到一个 8×8 的预测块输出, 其中输出的通道数表示的就是积极、消极的情感预测结果. 将这个在区间 [0, 1] 上的预测结果映射到区间 [0, 255], 并将消极情绪的预测值对应到 RGB 中的 R 通道, 将积极情绪的预测值对应到 RGB 中的 G 通道, B 通道全部初始化为 0. 然后采用临近插值法将此预测块插值到与原图大小一致, 最终按等权重对原图和预测结果进行融合. 部分结果见图 5.

从结果可以看出, 模型能够对这些图像做出基本的判断, 如图 5 第 1, 2, 4 幅画中对于面部把控和第 1, 2, 5 幅画中对于整体姿态的把控都基本符合尝试, 而在绘画图像的情感预测任务上还达不到像素级别的分类. 预测块以预测的区域的形式对原图像进行预测而不是单独的对每个像素进行识别, 因

为在训练过程中,由于找不到像素级别的标记,只能对原图像作为简单的二值分类任务进行预测,之后再对模型的输出采用邻近插值的方式变换为 448×448 的大小.因此,模型基本还是从整体上对图像进行预测而不是像素级别的预测,这也解释了预测图从整体上的把控是比较符合预期这一结论.图中结果也可以看出模型对于细粒度的把控是远远不到位的,但也展示了模型学习的基本情况.未来,我们将研究如何使模型结合一些更为细粒度的特征.

5 结语

针对云南省民族艺术绘画的情感分类问题,即绘画情感由大众化情感表示,本文设置了一系列实验并对实验结果进行性能分析,对比了几种常用的过采样方式对卷积神经网络在图像情感任务上性能的影响,包括民族艺术绘画图像数据和普通图片数据.为了验证本文方法的有效性,在 Twitter image dataset 上与目前最先进水平的方法进行了性能比较,实验证明本文所采用模型的有效性.此外,本文提出了一种相关任务的预训练策略,以提升模型在民族艺术绘画上的性能,并通过可视化的结构分析了模型学习的具体情况,这对于指导绘画艺术有一定的促进作用.

未来,我们将继续研究深度学习在图像情感计算中的应用,并将拓展到分析普通图像甚至视频的计算机视觉任务当中,也会分析如何利用噪声图像而不受干扰的学习模型,同时也会考虑结合生物信号如脑电、眼动等做更深入的分析.

参考文献

- 1 Picard R W. Affective Computing. London: MIT Press, 1997
- 2 Pang B, Lee L. Opinion mining and sentiment analysis. *FNT Inf Retrieval*, 2008, 2: 1–135
- 3 Yang Y H, Chen H H. Machine recognition of music emotion: a review. *ACM Trans Intell Syst Technol*, 2012, 3: 40
- 4 Wang W N, He Q H. A survey on emotional semantic image re-trieval. In: *Proceedings of IEEE International Conference on Image Processing*, San Diego, 2008. 117–120
- 5 Joshi D, Datta R, Fedorovskaya E, et al. Aesthetics and emotions in images. *IEEE Signal Process Mag*, 2011, 28: 94–115
- 6 Wang S F, Ji Q. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Trans Affect Comput*, 2015, 6: 410–430
- 7 Lee J, Park E J. Fuzzy similarity-based emotional classification of color images. *IEEE Trans Multimedia*, 2011, 13: 1031–1039
- 8 Lu X, Suryanarayan P, Adams R B, et al. On shape and the computability of emotions. In: *Proceedings of ACM International Conference on Multimedia*, 2012. 229–238
- 9 Machajdik J, Hanbury A. Affective image classification using features inspired by psychology and art theory. In: *Proceedings of ACM International Conference on Multi-media*, Firenze, 2010. 83–92
- 10 Solli M, Lenz R. Color based bags-of-emotions. In: *Proceedings of International Conference on Computer Analysis of Images and Patterns*, Munster, 2009. 573–580
- 11 Zhao S C, Gao Y, Jiang X L, et al. Exploring principles-of-art features for image emotion recognition. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, 2014. 47–56
- 12 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*, 2013, 35: 1798–1828
- 13 You Q Z, Luo J B, Jin H L, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks. 2015. ArXiv:1509.06041
- 14 Campos V, Jou B, Giró-i-Nieto X. From pixels to sentiment: fine-tuning CNNs for visual sentiment prediction. *Image Vision Comput*, 2017, 65: 15–22

- 15 Chang L, Chen Y F, Li F X, et al. Affective image classification using multi-scale emotion factorization features. In: Proceedings of International Conference on Virtual Reality and Visualization (ICVRV), 2016. 170–174
- 16 Rao T R, Xu M, Liu H Y, et al. Multi-scale blocks based image emotion classification using multiple instance learning. In: Proceedings of IEEE International Conference on Image Processing (ICIP), 2016. 634–638
- 17 Chen M, Zhang L, Allebach J P. Learning deep features for image emotion classification. In: Proceedings of IEEE International Conference on Image Processing, 2015. 4491–4495
- 18 You Q Z, Luo J B, Jin H L, et al. Building a large scale dataset for image emotion recognition: the fine print and the bench-mark. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016. 308–314
- 19 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1097–1105
- 20 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 21 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. 2014. arXiv:1409.4842
- 22 Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014. 580–587
- 23 Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector. In: Proceedings of European Conference on Computer Vision, 2015. 21–37
- 24 Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. 2016. ArXiv:1506.02640
- 25 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 3431–3440
- 26 Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets. 2014. ArXiv:1405.3531
- 27 Radenovi F, Tolias G, Chum O. CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples. In: Proceedings of European Conference on Computer Vision, 2016
- 28 Tajbakhsh N, Shin J Y, Gurudu S R, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imag, 2016, 35: 1299–1312
- 29 Jung H, Lee S, Yim J, et al. Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 2983–2991
- 30 Zhao B B. Design and implementation of emotion labeling system for Yunnan ethnic painting based on visual semantics. Dissertation for Master Degree. Kunming: Yunnan University, 2017 [赵贝贝. 基于可视化语义的云南民族绘画情感标注系统的设计与实现. 硕士学位论文. 昆明: 云南大学, 2017]
- 31 Ye X M, Cheng H. Color language and culture. Jiangxi Soc Sci, 2006, 3: 171–173 [叶小妹, 陈贺. 颜色语和文化. 江西社会科学, 2006, 3: 171–173]

Ethnic painting analysis based on deep learning

Hao ZHANG & Dan XU*

Department of Computer Science and Engineering, Yunnan University, Kunming 650504, China

* Corresponding author. E-mail: danxu@ynu.edu.cn

Abstract Besides conveying rich semantic information, images (especially artistic images) directly express peoples' emotions and influence the emotional levels of others. As the emotional reactions to different visual stimuli differ among individuals, the emotions contained in images must be properly understood. Taking ethnic painting image data as the research objects, this study analyzes how the hue, brightness, saturation and contrast of the paintings influences the classification results of convolutional neural networks. The analysis is performed by a fine-tuning method. When experimentally evaluated on a Twitter image dataset, the absolute accuracy of our method was 3.4% higher than a previous state-of-the-art method. Finally, we proposed pre-training strategies for related tasks, which significantly improve the emotion classification of ethnic paintings, and experimentally evaluated them on visualization structures.

Keywords ethnic painting, image emotion, convolutional neural network, fine-tuning, visualization

Hao ZHANG was born in 1993. Currently, he is a postgraduate student of Yunnan University. His research focuses on image semantic segmentation and image emotion semantic analysis.

Dan XU was born in 1968, and received her Ph.D. degree in 1999 from Zhejiang University. She is a doctoral supervisor and a member of CCF. Her main research areas are image analysis and understanding, visual calculation, and cognition.