软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2494−2506 [doi: 10.13328/j.cnki.jos.006487]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于深度多任务学习的图像美感与情感联合预测研究*

申 朕[1], 崔超然[2], 董桂鑫[2], 余 俊[3], 黄 瑾[1], 尹义龙[1]

[1](山东大学 软件学院, 山东 济南 250101)

[2](山东财经大学 计算机科学与技术学院, 山东 济南 250014)

[3](Department of Computer Science and Engineering, Lehigh University, Bethlehem PA 18015, USA)

通信作者: 崔超然, E-mail: crcui@sdufe.edu.cn; 尹义龙, E-mail: ylyin@sdu.edu.cn

摘 要: 图像美学评价和情感分析任务旨在使计算机可以辨认人类由受到图像视觉刺激而产生的审美和情感反应. 现有研究通常将它们当作两个相互独立的任务. 但是, 人类的美感与情感反应并不是孤立出现的; 相反, 在心理认知层面上, 两种感受的出现应是相互关联和相互影响的. 受此启发, 采用深度多任务学习方法在统一的框架下处理图像美学评价和情感分析任务, 深入探索两个任务间的内在关联. 具体来说, 提出一种自适应特征交互模块将两个单任务的基干网络进行关联, 以完成图像美学评价和情感分析任务的联合预测. 该模块中引入了一种特征动态交互机制, 可以根据任务间的特征依赖关系自适应地决定任务间需要进行特征交互的程度. 在多任务网络结构的参数更新过程中, 根据美学评价与情感分析任务的学习复杂度和收敛速度等差异, 提出一种任务间梯度平衡策略, 以保证各个任务可以在联合预测的框架下平衡学习. 此外, 构建了一个大规模的图像美学情感联合数据集 UAE. 据已有研究, 该数据集是首个同时包含美感和情感标签的图像集合. 本模型代码以及 UAE 数据集已经公布在 https://github.com/zhenshen-mla/Aesthetic-Emotion-Dataset.

关键词: 图像美学评价; 图像情感分析; 深度多任务学习; 自适应特征交互模块; 任务间梯度平衡策略

中图法分类号: TP391

中文引用格式: 申朕, 崔超然, 董桂鑫, 余俊, 黄瑾, 尹义龙. 基于深度多任务学习的图像美感与情感联合预测研究. 软件学报, 2023, 34(5): 2494–2506. http://www.jos.org.cn/1000-9825/6487.htm

英文引用格式: Shen Z, Cui CR, Dong GX, Yu J, Huang J, Yin YL. Unified Image Aesthetic and Emotional Prediction Based on Deep Multi-task Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2494–2506 (in Chinese). http://www.jos.org.cn/1000-9825/6487.htm

Unified Image Aesthetic and Emotional Prediction Based on Deep Multi-task Learning

SHEN Zhen[1], CUI Chao-Ran[2], DONG Gui-Xin[2], YU Jun[3], HUANG Jin[1], YIN Yi-Long[1]

[1](School of Software, Shandong University, Jinan 250101, China)

[2](School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China)

[3](Department of Computer Science and Engineering, Lehigh University, Bethlehem PA 18015, USA)

Abstract: Image aesthetic assessment and emotional analysis aim to enable computers to identify the aesthetic and emotional responses of human beings caused by visual stimulations, respectively. Existing research usually treats them as two independent tasks. However, people's aesthetic and emotional responses do not appear in isolation. On the contrary, from the perspective of psychological cognition, the two responses are interrelated and mutually influenced. Therefore, this study follows the idea of deep multi-task learning to deal with image aesthetic assessment and emotional analysis under a unified framework and explore their relationship. Specifically, a novel adaptive feature interaction module is proposed to correlate the backbone networks of the two tasks and achieve a unified prediction. In addition, a

* 基金项目: 国家自然科学基金 (61701281, 61876098); 国家重点研发计划 (2018YFC0830100, 2018YFC0830102); 山东省高等学校优势学科和人才团队培育计划
本文由"面向开放场景的鲁棒机器学习"专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.
收稿时间: 2021-05-28; 修改时间: 2021-07-16; 采用时间: 2021-09-27; jos 在线出版时间: 2022-10-14
CNKI 网络首发时间: 2022-11-15

© 中国科学院软件研究所 http://www.jos.org.cn

dynamic feature interaction mechanism is introduced to adaptively determine the degree of feature interaction between the tasks according to the feature dependencies. As the multi-task network updates structural parameters, the study, based on the inconsistency in complexity and convergence speed between the two tasks, proposes a novel gradient balancing strategy to ensure that the network parameters of each task can be smoothly learned under the unified prediction framework. Furthermore, the study constructs a large-scale unified image aesthetic and emotional dataset–UAE. According to the study, UAE is the first image collection containing both aesthetic and emotional labels. Finally, the model and codes of the proposed method as well as the UAE dataset have been released at https://github.com/zhenshen-mla/Aesthetic-Emotion-Dataset.

Key words: aesthetic assessment; emotion analysis; deep multi-task learning; adaptive feature interaction; gradient balancing strategy

伴随计算机视觉技术的快速发展, 人们不仅希望计算机能够在语义层面对图像的内容进行分析[1−3], 更期望计算机能够模拟人类视觉及思维系统, 产生更高层次的图像感知能力[4,5]. 作为感知理解研究中的两项代表性任务, 图像的美学评价[6]和情感分析[7]分别旨在使计算机可以辨认人类由受到图像视觉刺激而产生的审美和情感反应. 目前, 图像的美学评价和情感分析技术已经被应用在图像的检索[8]、管理[9]和增强[10]等方面. 例如, 在图像检索系统中, 考虑返回图像的情感倾向, 为用户提供语义准确且更有感染力的检索结果; 针对用户拍摄的关于同一物体或场景的多张候选照片, 筛选出最具美感的作品进行保存和展示, 合理地降低数据的存储开销; 在图像作品的创作和编辑中, 分析对比候选方案的美学质量, 提升作品的视觉美感. 通常情况下, 图像美学评价任务被形式化为一个区分高美感与低美感图像的分类或回归问题[11,12], 而图像情感分析则被作为一个将图像划分为预定义情感类别的问题[13]. 在美学评价和情感分析的研究过程中, 早期方法主要关注于如何手工地设计有效的图像特征. 近年来, 随着深度学习技术的快速发展, 现有的图像美学评价和情感分析研究大多利用卷积神经网络自动地提取具有良好区分能力的图像特征[14,15], 并实现了性能的大幅提升. 值得注意的是, 现有技术通常将图像的美学评价和情感分析当作两个相互独立的任务. 但直觉上, 人类的美感与情感感受并不是孤立出现的; 相反, 在心理认知层面上, 两种感受的出现应是相互关联和相互影响的. 例如, 如果一幅图像能够使人类获得审美上的愉悦, 那它也很有可能会唤起观察者的积极情感. 神经科学领域的研究[16]也表明, 人类的审美体验是一种伴随着情感状态不断升级的认知过程, 反之亦然. 因此, 本文认为图像美学评价和情感分析应是两个紧密关联的视觉理解任务. 但目前, 图像美学评价与情感分析任务间的关系尚未得到充分地探讨与研究.

受此启发, 本文提出基于多任务学习的思想来同时解决图像美学评价和情感分析问题, 使用多任务卷积神经网络来利用两个任务间潜在的相关性, 进而同时实现对两个任务的性能提升[17]. 在实现多任务卷积神经网络时, 比较常用的是参数硬共享策略, 即不同任务共享较低的网络层, 并在较高的网络层上维持各自任务的分支. 但这种做法需要预先人工指定网络共享层和分离层的位置, 对于共享与分离位置不合理的选择可能会导致方法性能的严重下滑[18]. 为了解决该问题, 最近的研究多采用参数软共享的多任务学习方案, 即学习如何在单个网络层上有效地融合不同任务的特征, 从而实现不同任务间的知识共享[19]. 例如, Misra 等人[18]提出一个十字绣 (cross stitch) 网络结构来学习不同任务对应通道位置特征的线性组合, 并将组合后的特征输入到对应任务分支的下层网络中. Gao 等人[20]提出了一个分层特征融合网络, 将来自不同任务分支的特征图首先沿通道维度进行堆叠, 然后经过 1×1 卷积进行降维处理, 以满足下一层的维度要求.

尽管上述研究取得了很大的进展, 但它们在模型训练结束之后, 对每个任务的特征采取固定参数的组合方案, 而不考虑特征信息本身的特性. 事实上, 不同任务所提取的特征存在一定程度的差异性. 为了进一步说明这一点, 本文使用两个 ResNet50 分类网络[1]作为多任务结构的两个分支, 并使用该结构对美学评价与情感分析任务展开训练. 训练结束后, 本文将数据集中随机挑选的若干图像输入到多任务结构中. 图 1 展示了图像在多任务结构中两个分支对应位置处的特征图, 分别使用 ResNet50 结构进行图像美学评价和情感分析. 对每张输入图像, 可视化两个任务分支对应位置处的特征图. 可以看到, 左边 3 列的图像, 其特征图非常相似, 而右边 3 列图像的特征图差异很大甚至互补. 直觉上, 一个合适的多任务学习方法应该能够在前者情况下保留更多各自任务的特征, 而在后者情况下可以选择更广泛的共享表示以完成从一个任务传递有效特征到另一个任务. 因此, 在多任务结构中维持一个固定参数的特征组合策略可能是不合适的, 需要考虑各任务特征信息的特性. 受此启发, 本文提出了一种新颖的自适应特征交互 (adaptive feature interaction, AFI) 模块将多个单任务的基干网络进行关联, 并根据各任务输入特征

© 中国科学院软件研究所 http://www.jos.org.cn

信息的特性自适应地决定任务间特征的共享程度.

多任务学习的另一个关键挑战是在模型训练时如何平衡不同的任务[21,22]. 实际中, 不同任务具有不同的复杂度和收敛速度. 如果在没有任何平衡控制的情况下对多任务学习网络结构进行训练, 训练过程中的反向传播算法很容易被某个任务的梯度所主导, 从而降低其他任务的性能. 针对该问题, 本文为图像美学与情感的多任务学习结构提出了一种新颖的任务间梯度平衡策略: 梯度重校准, 在参数软共享的多任务学习结构中对特征交互模块回传给各网络分支的梯度进行重校准. 具体来说, 以各网络分支的梯度量级为标准, 对特征交互模块回传的梯度进行缩放, 并将各分支梯度和特征交互模块梯度进行整合, 以保证两个任务在统一学习的框架下能够平衡学习.

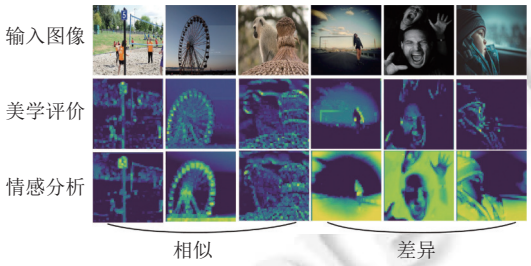

图 1 图像在多任务结构中两个分支对应位置处的特征图

除此之外, 目前业界缺乏同时带有美学标记和情感标记的大规模图像数据集. 为了顺利开展图像美学情感的多任务研究, 本文创建了一个大规模美学情感联合数据集 UAE (unified aesthetic and emotional dataset). 据本文所知, 该数据集是第 1 个将美感与情感标签相关联的图像集合, 为本文研究美感与情感任务的联合预测奠定了基础. 目前, 该数据集已经公开发布供他人研究和使用 (https://github.com/zhenshen-mla/Aesthetic-Emotion-Dataset).

实验中, 本文对所提出的方法进行了详细的消融研究. 并分别与典型的图像美学评价方法、图像情感分析方法以及多任务学习方法进行对比. 实验结果表明, 本文方法优于目前最先进的方法. 此外, 通过与近期提出的其他多任务平衡方法进行比较, 本文证明了所提出的梯度重校准策略可以更好地控制图像美学情感联合学习的训练过程.

本文的贡献主要包括 3 个方面.

(1) 本文提出了一种新颖的自适应特征交互模块, 利用该模块连接多任务深度网络中不同任务分支的中间卷积层, 并根据输入特征信息的特性进行任务间特征的自适应交互, 形成更有效的共享特征表示.

(2) 本文提出了一种新颖的多任务间梯度平衡策略, 对基于参数软共享的多任务结构中特征交互模块所回传的梯度进行重校准, 以保证不同任务在统一的框架下能够平衡学习. 值得注意的是, 现有研究均是面向基于参数硬共享的多任务网络结构.

(3) 本文创建了一个大规模图像美学情感联合数据集 UAE, 该数据集是第 1 个将美感与情感标签相关联的图像集合, 为本文研究美感与情感任务的联合预测奠定了基础. 该数据集已经公布给科研界以供学习和研究使用.

本文第 1 节介绍图像美学评价、情感分析以及多任务学习等方面的相关工作及国内外研究现状. 第 2 节描述本文创建 UAE 数据集的具体过程, 并对数据集进行了统计分析. 第 3 节介绍了基于自适应特征交互的多任务学习方法以及参数更新时任务间梯度平衡策略的具体实现. 第 4 节进行了本文方法的消融研究, 以明确本文方法的实现细节和具体作用. 第 5 节详细展示了实验对比结果及分析. 第 6 节进行了可视化研究, 更直观地理解本文方法的效果. 第 7 节对本文进行总结并讨论了未来工作的方向.

1 研究现状

在本节中, 本文简要回顾之前研究中典型的图像美学评价与情感分析方法. 然后介绍一下目前经典的多任务神经网络结构与任务间梯度平衡策略.

1.1 图像美学评价

近年来, 由于其潜在的应用, 图像美学评价取得了广泛关注. 早期的研究大都是基于人类的审美直觉或摄影规

© 中国科学院软件研究所 http://www.jos.org.cn

则来手工制作特征. 随着深度学习的兴起, 卷积神经网络被广泛应用于自动学习美学特征, 并取得了最先进的性能. 例如, Lu 等人[15]提出了一种特征学习以及聚集不同 patch 特征的框架, 尽量避免丢失细纹理特征. Datta 等人[23]通过设计多类视觉特征, 如色彩指标、三分法、景深等, 来区分美观与不美观的图像. Dhar 等人[24]从图像的布局、内容、光照等方面提取了一些高层次的可描述属性来预测感知审美质量. Tang 等人[25]认为不同类型的图像与不同的审美评价标准有关, 并根据图像内容的多样性, 以不同的方式设计特征. Marchesotti 等人[26]使用通用图像描述符 (包括视觉词袋和 Fisher 向量) 来评估美学质量, 并显示出比传统手工制作特征更好的性能. 随着全卷积神经网络的提出, Fang 等人[27]提出了利用全卷积神经网络来解决图像输入尺寸限制问题的图像美学质量评价模型. Kao 等人[11]通过将语义识别作为相关任务来辅助美学评价, 提出了多任务卷积神经网络框架.

1.2 图像情感分析

传统的图像情感分析方法主要聚焦于手工提取图像视觉情感特征. 特征提取需要从人类的心理、生理特点出发, 选取和人的情感、情绪密切相关的视觉特征, 并选择合适的描述方式. 在目前的情感研究中选择最多的特征包括颜色、纹理、形状和轮廓等[28−30]. 在近期的研究中, 已经有研究人员开始将深度学习模型应用到视觉情感研究中. Chen 等人[13]提出了一种基于卷积神经网络进行视觉情感概念分类的方法. 在网络模型上, 他们采用了一种对大规模图像数据集 ImageNet[31]进行分类时表现优异的深层卷积神经网络. You 等人[32]在利用深度卷积网络进行情感预测的同时, 提出了一种渐进微调的方案, 即在初次训练完成后, 对训练数据样本进行反向筛选, 排除干扰噪声数据, 然后在原来的基础上用筛选后的样本进行优化调整训练.

1.3 深度多任务学习

多任务学习方法是机器学习领域的一种学习范式, 基于任务间的共享表示, 把多个相关任务放在同一模型中进行联合学习. 最近, 出现了许多关于多任务神经网络结构的研究, 并取得了令人印象深刻的进展. Misra 等人[18]提出了十字绣结构来学习不同任务对应通道位置特征的线性组合, 并将组合后的特征输入到对应任务分支的邻接层中. Gao 等人[20]提出了分层特征融合网络, 来自不同任务分支的特征图首先沿通道维度进行堆叠, 然后经过 1×1 卷积进行降维处理, 以满足下一层的维度要求. 此外, Kawakami 等人[33]通过 1×1 卷积将单任务网络的特征映射连接起来, 并在任务间使用不同数据集进行多任务学习. Liu 等人[34]提出一种多任务注意力网络, 该网络由一个单一的共享网络和用于各任务的软注意力模块组成, 由此可以学习共享和特定任务的特征信息.

目前, 多数关于多任务学习的研究都聚焦于网络结构的设计. 然而, 对于多任务学习损失的优化也非常重要. 在传统的硬参数共享的结构中, 所有任务共享输入层和网络浅层, 然后每个任务的顶层分别拟合各自输出. 对于多任务的损失, 最简单的方式是直接将所有任务的损失进行相加, 得到整体的损失. 然而, 不同任务损失的量级很可能不一样, 损失直接相加的方式有可能会导致多任务学习被某个任务所控制. 当模型倾向于去拟合某个任务时, 其他任务往往会受到负面影响, 效果相对变差. 针对该问题, Chen 等人[35]提出一种梯度标准化策略, 使不同任务的损失量级接近, 并以相近的收敛速度进行学习. Kendall 等人[36]基于偶然不确定性中的任务依赖型、同方差不确定性来进行建模, 使不确定性大的任务权重变小, 噪声小且简单的任务权重会变大. Liu 等人[34]使用了各任务损失的动态加权平均策略, 将损失下降速度快的任务赋予较小的权重, 反之权重会变大, 从而使各任务以相对平衡的方式进行参数更新.

不同于上述方法, 本文使用自适应特征交互模块将多个单任务的基干网络进行关联, 并根据通道间特征依赖关系进行建模, 以推断不同任务的交互权值, 从而实现图像美学与情感任务间特征的自适应交互. 并使用梯度重校准策略对参数软共享的多任务训练过程进行约束, 将自适应特征交互模块回传给美学和情感网络分支的梯度进行重校准, 从而实现各任务平衡学习.

2 UAE 数据集

在本节中, 本文构建了一个图像美学情感联合数据集, 即 UAE 数据集 (unified aesthetic and emotional dataset), 为本文研究美感与情感任务的联合预测奠定了基础. 在 UAE 数据集创建之初, 本文对现有的涉及大规模高层次视觉感知任务的图像集合做了充分调查与研究, 发现 You 等人[37]提出的 FI 数据集内容丰富, 包含准确多样的图像

© 中国科学院软件研究所 http://www.jos.org.cn

情感标签信息, 且数据样本符合现实场景. 因此, 本文在 FI 数据集的基础上, 进一步扩展了图像的美感标签, 以满足图像美学与情感任务的联合预测.

具体来说, FI 数据集中的图像主要从互联网图像搜索引擎 (Flickr 和 Instagram) 获得, You 等人[37]首先通过使用 8 种情绪词 (娱乐, 愤怒, 敬畏, 满足, 厌恶, 兴奋, 恐惧, 悲伤) 作为查询关键字来收集体现情感内容的图像. 然后, 对于收集到的图像, 首先由众包网站上的志愿者对其进行打分, 初步获得带有“弱情感标记”的图像集合. 进一步地, 通过资格测试对志愿者进行筛选, 其中资格测试共有 20 道关于图像情感的题目, 如“这幅图像让你感到愤怒吗”回答“是”或“否”, 至少答对 10 道题目才能成为专家. 最后, 邀请专家对图像进行重新打分, 基于专家的打分结果, 选择出带有“强情感标记”的图像集合. FI 数据集总共包含约 23 000 幅图像, 共 8 种情感类别, 每种情感类别的图像数量均大于 1 100 张.

在标注图像的美感标签时, 出于准确性和便捷性的综合考虑, 本文为每幅图像的美感程度给出了 4 种选择, 即完美 (10 分), 良好 (7 分), 一般 (4 分) 和丑陋 (1 分). 对于待评价的图像, 志愿者需要根据自身审美选择其中一种美感程度. 在评分开始时, 本文随机地从 FI 数据集中选择了 800 幅图像 (每个情感类别 100 幅图像), 并邀请 20 名志愿者 (男性一半, 女性一半) 对它们进行美感打分. 根据每位志愿者对 800 幅图像的打分结果 (800 维向量), 本文计算了任意两个志愿者之间打分分布的 KL 散度以求分布间的差异性, 然后筛选掉 10 名美学偏差较大的志愿者. 然后, 本文将剩余的 10 位志愿者作为专家来对数据库中的全部图像进行美感打分. 评分结束后, 计算每幅图像获得的平均得分. 最后, 本文通过设置中间阈值的方法来划分已标记好的数据集, 其中阈值设置为 5, 即美感评分大于 5 为高美感图像, 否则为低美感图像, 两个类别的图像数量分别为 12 641 张和 9 445 张. 接下来, 本文将对 UAE 数据集中的全部数据进行统计分析.

UAE 数据集中图像的美感与情感类别分布如图 2 所示. 可以看出, 具有正向情感 (娱乐, 兴奋, 敬畏, 满足) 的图像样本更有可能得到较高的美学评价, 而具有负面情感 (厌恶, 愤怒, 恐惧, 悲伤) 的图像样本更倾向于获得较低的美学评价. 这样的统计结果也验证了本文的猜想, 即图像美学评价与情感分析任务之间存在较强的相关性, 应该将两个任务协同处理. 图 3 进一步展示了 UAE 数据集中的部分示例图像.

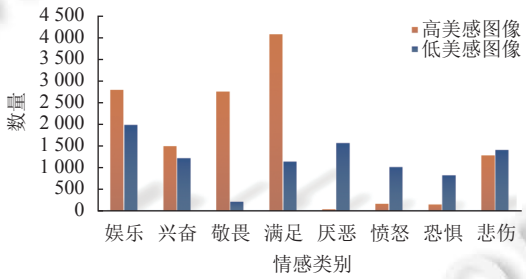

图 2 图像统计直方图

图 3 UAE 数据集示例图像

3 方 法

最近, 基于深度神经网络的多任务学习模型受到广泛关注. 本文提出了一种面向图像美学评价和情感分析联合预测的多任务卷积神经网络. 在本节中, 首先阐述自适应特征交互模块的总体结构, 重点描述任务间特征交互的设计方案, 最后讨论关于美学情感多任务学习中的任务间梯度平衡策略.

© 中国科学院软件研究所 http://www.jos.org.cn

3.1 自适应特征交互模块

在联合学习的总体架构中, 自适应特征交互 (adaptive feature interaction, AFI) 模块被用来连接美学评价任务 T_A 和情感分析任务 T_E 网络分支对应位置的卷积层. 但当 T_A 和 T_E 具有不同的网络结构时, AFI 模块则用来连接两个任务对应作用的卷积层. 如图 4 所示, T_A 和 T_E 保持各自的网络结构不变, 通过 AFI 模块进行通信, 形成基于参数软共享[17]的多任务学习结构. AFI 模块的结构如图 4. 其中, $\boldsymbol{x}_A \in R^{H\times W\times C}$ 和 $\boldsymbol{x}_E \in R^{H\times W\times C}$ 分别是 T_A 和 T_E 网络分支在某一对应位置卷积层所输出的特征图, 其中 H, W 和 C 分别代表特征图的长度, 宽度和通道数. AFI 模块接收 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 作为输入, 并将它们的对应元素相加以完成特征交互:

$$\hat{\boldsymbol{x}}_A = \boldsymbol{w}_{\text{self}} \times \boldsymbol{x}_A + \boldsymbol{w}_{\text{other}} \times \boldsymbol{x}_E, \quad \hat{\boldsymbol{x}}_E = \boldsymbol{w}_{\text{self}} \times \boldsymbol{x}_E + \boldsymbol{w}_{\text{other}} \times \boldsymbol{x}_A \tag{1}$$

其中, $\hat{\boldsymbol{x}}_A$ 和 $\hat{\boldsymbol{x}}_E$ 是 AFI 模块的输出, $\boldsymbol{w}_{\text{self}}$ 和 $\boldsymbol{w}_{\text{other}}$ 表示 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 在进行特征交互时的权重. 直观上说, $\boldsymbol{w}_{\text{self}}$ 指的是 $\boldsymbol{x}_A$ 或 $\boldsymbol{x}_E$ 在特征交互过程中保持自身特征信息的程度, $\boldsymbol{w}_{\text{other}}$ 表示从其他任务中获取特征信息的程度. 为了调整 $\hat{\boldsymbol{x}}_A$ 和 $\hat{\boldsymbol{x}}_E$ 在特征交互后能处在一个合理的数值范围内, 本文要求 $\boldsymbol{w}_{\text{self}}$ 和 $\boldsymbol{w}_{\text{other}}$ 对应元素和为 1, 即:

$$\boldsymbol{w}_{\text{self}} + \boldsymbol{w}_{\text{other}} = 1 \tag{2}$$

本文寻求一个特征动态交互机制, 通过考虑 T_A 和 T_E 之间的通道依赖性, 允许 T_A 和 T_E 根据自身输入 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 来动态地调整交互权重 $\boldsymbol{w}_{\text{self}}$ 和 $\boldsymbol{w}_{\text{other}}$. 在第 3.1 节中, 本文将详细介绍 AFI 模块中的特征交互方案.

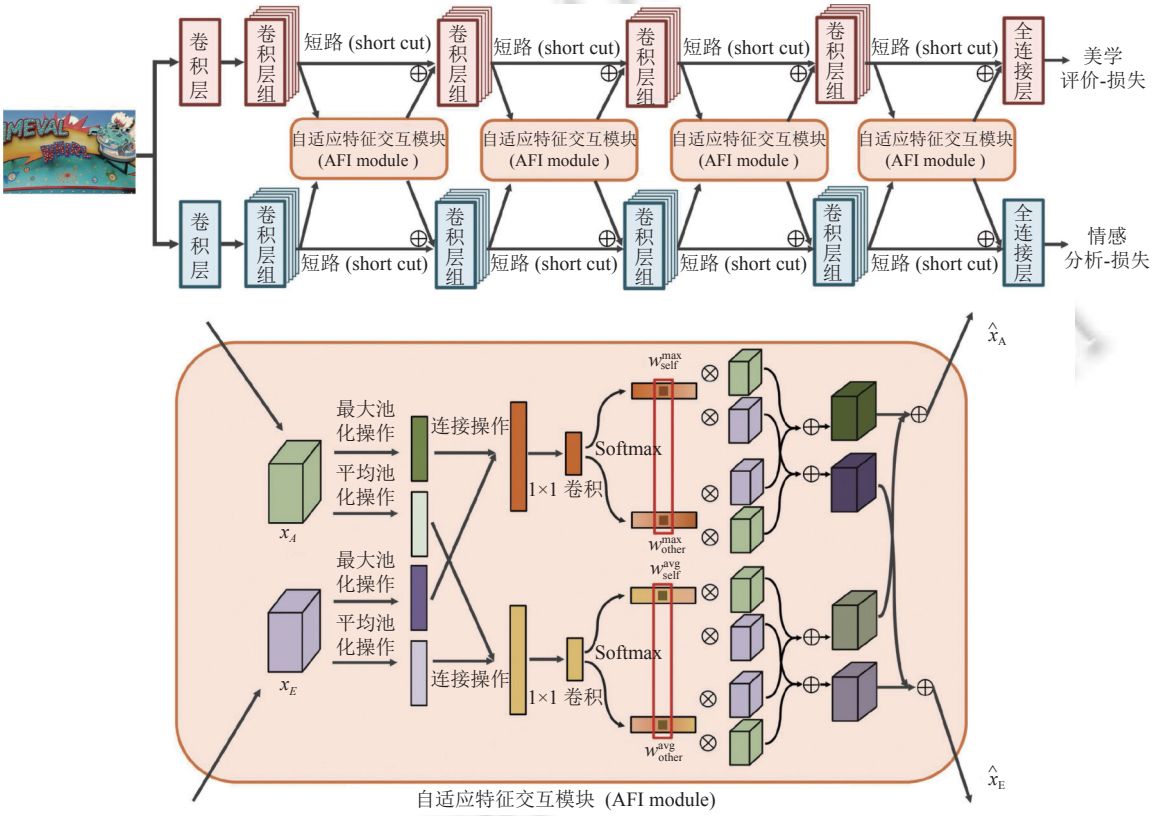

图 4 自适应特征交互模块结构图

最终, 将交互后的特征与每个任务的原有特征相加, 即:

$$\boldsymbol{x}'_A = \hat{\boldsymbol{x}}_A + \boldsymbol{x}_A, \quad \boldsymbol{x}'_E = \hat{\boldsymbol{x}}_E + \boldsymbol{x}_E \tag{3}$$

$\boldsymbol{x}'_A$ 和 $\boldsymbol{x}'_E$ 将分别被传递到 T_A 和 T_E 网络分支的下一个卷积层. 这里也体现了残差学习[1]的思想, AFI 模块用于连接美学情感网络分支的中间层, 根据输入特征特性进行自适应的特征交互处理. 随之短路 (short cut) 的特征与

© 中国科学院软件研究所 http://www.jos.org.cn

AFI 模块的输出特征进行加和, 并输入到网络下层, 从而完成残差学习. 在如今网络层数极深的情况下, 以上方式可以减轻深度网络存在的设计难度高和收敛速度慢的问题. 并缓解随着网络层数增多, 拟合效果反而变差的问题.

3.2 特征交互方案

在任务间进行特征交互的过程中, 为了有效地获取各自任务的特征描述符, AFI 模块首先沿 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 的空间维度进行压缩. 通过在空间维度对各自任务特征图进行全局平均池化操作和最大池化操作以得到通道上下文信息描述符 $\boldsymbol{d}_A^{\max} \in R^{C\times 1}$, $\boldsymbol{d}_A^{\text{avg}} \in R^{C\times 1}$, $\boldsymbol{d}_E^{\max} \in R^{C\times 1}$, $\boldsymbol{d}_E^{\text{avg}} \in R^{C\times 1}$. 在之前的研究中[38], 通常单独的使用某一种池化策略, 但本文认为两种池化策略在作用上可以相互补充, 从而形成更全面的通道上下文信息描述符[39]. 因此, 在本文中联合使用最大池化和平均池化操作:

$$\begin{cases} \boldsymbol{d}_A^{\max}(k) = \max_{(i,j)}(\boldsymbol{x}_A(i,j,k)), \quad \boldsymbol{d}_A^{\text{avg}}(k) = \dfrac{1}{H\times W}\displaystyle\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_A(i,j,k) \\ \boldsymbol{d}_E^{\max}(k) = \max_{(i,j)}(\boldsymbol{x}_E(i,j,k)), \quad \boldsymbol{d}_E^{\text{avg}}(k) = \dfrac{1}{H\times W}\displaystyle\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_E(i,j,k) \end{cases} \tag{4}$$

其中, $\boldsymbol{d}_A^{\max}(k)$ 是 $\boldsymbol{d}_A^{\max}$ 的第 k 元素, $\boldsymbol{x}_A(i,j,k)$ 指的是 $\boldsymbol{x}_A$ 的第 k 层特征图中位置在 (i,j) 的元素值, 其他同理. 然后, 本文分别对 $\boldsymbol{d}_A^{\max}(k)$ 和 $\boldsymbol{d}_E^{\max}(k)$, $\boldsymbol{d}_A^{\text{avg}}(k)$ 和 $\boldsymbol{d}_E^{\text{avg}}(k)$ 进行连接操作, 并将其输入到具有 *ReLU* 激活函数的 1×1 卷积层中进行降维处理. 通过以上方式对 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 的通道描述符进行编码, 以捕获两者之间的依赖关系:

$$\boldsymbol{l}^{\max} = ReLU(conv_{1\times 1}([\boldsymbol{d}_A^{\max}, \boldsymbol{d}_E^{\max}])), \quad \boldsymbol{l}^{\text{avg}} = ReLU(conv_{1\times 1}([\boldsymbol{d}_A^{\text{avg}}, \boldsymbol{d}_E^{\text{avg}}])) \tag{5}$$

其中, $conv_{1\times 1}$ 表示卷积核尺寸为 1×1 的卷积层, 并产生输出 $\boldsymbol{l}^{\max} \in R^{\frac{2C}{r}\times 1}$ 和 $\boldsymbol{l}^{\text{avg}} \in R^{\frac{2C}{r}\times 1}$. 在对通道描述符进行编码操作时, 不仅限于使用 *ReLU* 激活函数与 1×1 卷积层的方式, 同样可以采用其他方式如全连接层与激活函数的搭配, 本文选取了其中一种方式. 遵循之前的研究[40], 本文将 r 设为 16.

接下来, 本文将 $\boldsymbol{l}^{\max}$ 和 $\boldsymbol{l}^{\text{avg}}$ 作为引导产生 T_A 和 T_E 的特征交互权重 $\boldsymbol{w}_{\text{self}} \in R^{C\times 1}$ 和 $\boldsymbol{w}_{\text{other}} \in R^{C\times 1}$, 并将 Softmax 算子应用于 $\boldsymbol{w}_{\text{self}}$ 和 $\boldsymbol{w}_{\text{other}}$ 的对应位置元素, 因此满足公式 (2) 的约束:

$$\begin{cases} w_{\text{self}}^{\max}(k) = \dfrac{e^{U^{\max}(k,\cdot)l^{\max}}}{e^{U^{\max}(k,\cdot)l^{\max}} + e^{V^{\max}(k,\cdot)l^{\max}}}, \quad w_{\text{other}}^{\max}(k) = \dfrac{e^{V^{\max}(k,\cdot)l^{\max}}}{e^{U^{\max}(k,\cdot)l^{\max}} + e^{V^{\max}(k,\cdot)l^{\max}}} \\ w_{\text{self}}^{\text{avg}}(k) = \dfrac{e^{U^{\text{avg}}(k,\cdot)l^{\text{avg}}}}{e^{U^{\text{avg}}(k,\cdot)l^{\text{avg}}} + e^{V^{\text{avg}}(k,\cdot)l^{\text{avg}}}}, \quad w_{\text{other}}^{\text{avg}}(k) = \dfrac{e^{V^{\text{avg}}(k,\cdot)l^{\text{avg}}}}{e^{U^{\text{avg}}(k,\cdot)l^{\text{avg}}} + e^{V^{\text{avg}}(k,\cdot)l^{\text{avg}}}} \end{cases} \tag{6}$$

其中, $\boldsymbol{U}^{\max} \in R^{C\times\frac{2C}{r}}$, $\boldsymbol{V}^{\max} \in R^{C\times\frac{2C}{r}}$ 是参数矩阵, 用来将引导符 $\boldsymbol{l}^{\max}$ 转换为 $\boldsymbol{w}_{\text{self}}^{\max}$ 和 $\boldsymbol{w}_{\text{other}}^{\max}$. $U^{\max}(k,\cdot)$ 和 $V^{\max}(k,\cdot)$ 分别表示其第 k 行, 其他同理. 直观上, 以上的操作相当于将软注意力机制[39]应用于 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 的通道维度上. 最后, 特征交互权重 $\boldsymbol{w}_{\text{self}}$ 和 $\boldsymbol{w}_{\text{other}}$ 沿 $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_E$ 的空间维度展开, 通过公式 (1) 来实现特征的自适应交互.

3.3 任务间梯度平衡策略

多任务学习的另一个关键挑战是在模型训练中如何平衡不同的任务. 在实验中, 本文注意到不同任务在梯度量级和收敛速度上有很大差异. 如果在没有任何平衡控制的情况下共同训练两个任务, 训练过程中的反向传播算法很容易被某个任务的梯度所主导, 从而影响其他任务的性能. 针对这一问题, 近期的研究也提出了几种多任务梯度平衡策略[34−36]. 但值得注意的是, 这些策略均面向参数硬共享的多任务学习方式. 在这些策略中, 根据不同任务训练难度和收敛速度的差异性建模, 核心问题在于如何确定各个任务损失的权重.

针对参数软共享的多任务学习方式, 本文提供了一种新颖的任务间梯度平衡策略: 梯度重校准. 实验中本文发现, 情感分析任务在训练过程中前期的梯度量级约是美学评价任务梯度量级的 10 倍左右, 在训练结束时大约是 2–3 倍. 在梯度反向传播时, 经过 AFI 模块后, 任务间的梯度彼此相互影响, 导致美学评价任务的梯度量级上升多倍, 偏离了原来的学习方向. 故在本文的研究中, 任务间梯度量级的不匹配是主要问题. 不同于之前研究中对各任务损失进行加权求和的方式, 本文方法是在模型梯度计算时对 AFI 模块所回传梯度进行重校准, 从而保证其能够与单个任务分支回传的梯度进行安全的整合. 以美学评价任务网络分支为例, 对于任一卷积层组, 指定 $\boldsymbol{g}_{\text{AFI}}^{A}$ 为经由 AFI 模块回传的梯度值. $\boldsymbol{g}_{\text{SC}}^{A}$ 为经由短路回传的梯度值. 由于 $\boldsymbol{g}_{\text{AFI}}^{A}$ 含有来自情感分析任务的大量级梯度, 在将 $\boldsymbol{g}_{\text{AFI}}^{A}$ 和 $\boldsymbol{g}_{\text{SC}}^{A}$ 整合前需要对 $\boldsymbol{g}_{\text{AFI}}^{A}$ 进行缩小处理:

© 中国科学院软件研究所 http://www.jos.org.cn

$$f(\boldsymbol{g}_{\mathrm{AFI}}^{A},\boldsymbol{g}_{\mathrm{SC}}^{A})=\begin{cases}T/(\log(\|\boldsymbol{g}_{\mathrm{AFI}}^{A}\|/\|\boldsymbol{g}_{\mathrm{SC}}^{A}\|)+1)\times\boldsymbol{g}_{\mathrm{AFI}}^{A}, & \|\boldsymbol{g}_{\mathrm{AFI}}^{A}\|/\|\boldsymbol{g}_{\mathrm{SC}}^{A}\|>1\\ 1\times\boldsymbol{g}_{\mathrm{AFI}}^{A}, & \|\boldsymbol{g}_{\mathrm{AFI}}^{A}\|/\|\boldsymbol{g}_{\mathrm{SC}}^{A}\|\leqslant 1\end{cases} \tag{7}$$

本文使用梯度的 l_2 范式来进行量级比较. 当$\|\boldsymbol{g}_{\mathrm{AFI}}^{A}\|$不大于$\|\boldsymbol{g}_{\mathrm{SC}}^{A}\|$时, 表明 AFI 模块的回传梯度与美学评价分支的自身梯度量级相当, 可以直接进行整合. 当$\|\boldsymbol{g}_{\mathrm{AFI}}^{A}\|$大于$\|\boldsymbol{g}_{\mathrm{SC}}^{A}\|$时, 则需要对其进行缩小处理. 其中 T 是通过网络自学习的参数[41], 初始化值为 1.0. 使用对数函数的目的是增加非线性, 与 T 协同控制对 AFI 模块回传梯度进行放缩的程度.

在对 AFI 模块所回传梯度进行重校准后, 可以与美学评价网络分支的回传梯度 $\boldsymbol{g}_{\mathrm{SC}}^{A}$ 进行安全的整合:

$$\boldsymbol{g}^{A}=f(\boldsymbol{g}_{\mathrm{AFI}}^{A},\boldsymbol{g}_{\mathrm{SC}}^{A})+\boldsymbol{g}_{\mathrm{SC}}^{A} \tag{8}$$

对于情感分析任务网络分支, 本文采取相同的梯度重校准策略.

4 消融实验

在本节中, 本文进行一系列的实验, 从不同的角度评估本文方法. 所有的实验均在配置有 10 核 2.4 GHz 的 Intel Xeon 处理器, NVIDIA Titan XP 显卡的工作站上进行. 对于美学评价与情感分析等图像级别的分类任务, 本文使用准确率 *Accuracy* 和 F_1 得分作为评价准则. 其中 F_1 得分是分类精度 *Precision* 和召回率 *Recall* 的调和平均值. 在所有后续的实验中如未做特殊说明, 都将使用 UAE 数据集中的全部数据进行实验, 并将全部的 22 086 幅图像随机分为训练集 (70%, 15 460 幅图像)、验证集 (10%, 2 209 幅图像) 和测试集 (20%, 4 417 幅图像).

对于自适应特征交互模块, 有 3 个主要因素影响本文方法的性能. 一是在基干网络中自适应特征交互模块的插入位置, 其次是网络的参数初始化策略, 最后是在特征交互时选择合适的池化方式. 本文使用 ResNet50 结构作为基干网络进行消融实验.

4.1 自适应特征交互模块插入位置

ResNet 网络共分为 4 个模块 (block)[1], 依次堆叠而成. ResNet 网络的浅层提取的主要是底层特征, 以目标轮廓、色彩以及具体位置有关的图像信息. 而在网络的深层, 其主要提取的是抽象且高维的语义信息. 不同层次的特征交互可能带来不一样的结果. 如果在浅层进行特征交互, 则偏重于细粒度信息, 任务间分享的是偏具体的概念. 而在深层进行交互, 任务间共享的则是高维的抽象信息. 所以, 在网络的浅层以及深层进行特征交互各有特点, 不同形式的特征交互可能产生不同的实验效果, 接下来进行消融研究来验证假设. 在实验中, 按网络层的深度, 将 ResNet 网络的前两个模块作为浅层模块, 后两个模块作为深层模块. 表 1 展示了特征交互模块插入位置的消融实验结果. 从结果来看, 本文发现使用特征交互模块会比没有交互的美学和情感单任务有明显的性能提升. 而且在网络浅层与深层均插入特征交互模块可以带来更明显的性能提升, 即任务间交互既需要浅层的细粒度特征信息又需要深层的高维抽象信息. 因此, 本文方法在基干网络的每个模块的后面均插入自适应特征交互模块, 这样既能维持基干网络的结构不变, 又能以最有效的方式进行特征交互.

表 1 自适应特征交互模块的插入位置对方法性能的影响 (%)

自适应特征交互模块插入位置	美学评价		情感分析	
	Accuracy	F_1	*Accuracy*	F_1
单任务基线	79.28	77.31	64.36	54.27
仅浅层	80.25	78.37	65.53	54.37
仅深层	80.34	78.30	65.35	54.31
浅层+深层	**80.55**	**78.61**	**65.85**	**54.90**

4.2 初始化策略

在对网络模型进行参数初始化时, 本文考虑了 3 种初始化策略. 第 1 种是传统的 Xavier 初始化方法[42]. 另外, 本文采用加载单任务网络预训练模型的方法, 这也是目前深度学习中最常采用的一种初始化方法. 具体来说, 本文分别引入在 ImageNet 数据集和 UAE 数据集上获得的预训练模型的权重作为多任务模型的初始化参数. 如后文表 2 展示了 3 种初始化策略的实验结果. 可以看到, 使用加载单任务网络预训练模型的初始化方法具有更好的性

© 中国科学院软件研究所 http://www.jos.org.cn

能, 在各个评价指标上均很大程度优于 Xavier 方法. 而在 UAE 数据集上的初始化策略又略微优于 ImageNet 数据集. 因此, 在目标单任务的网络模型训练结束之后, 部署本文方法更加实用.

4.3 池化策略

在特征交互过程中选择不同的池化策略, 会使生成的特征交互权重具有不同的关注点. 在进行选择时, 本文考虑了 3 种池化策略, 即平均池化操作、最大池化操作以及两者的组合操作. 不同池化操作对方法性能的影响如表 3 所示. 我们发现, 最大池化方式要比平均池化更适合本文的方法, 而最大池化与平均池化的组合达到了最好效果. 直观上解释, 综合利用两种池化操作可能会形成更全面的特征上下文信息描述符. 因此, 在本文中, 本文采用最大池化与平均池化相结合的方式.

表 2 网络模型初始化方式对方法性能的影响 (%)

初始化权重	美学评价		情感分析	
	Accuracy	F_1	*Accuracy*	F_1
Xavier	67.25	63.17	45.19	32.08
ImageNet	80.13	77.26	64.38	54.22
UAE	**80.55**	**78.61**	**65.85**	**54.90**

表 3 池化策略对方法性能的影响 (%)

池化策略	美学评价		情感分析	
	Accuracy	F_1	*Accuracy*	F_1
最大池化	80.01	77.22	64.62	54.17
平均池化	79.84	77.03	64.11	54.12
两种方式组合	**80.55**	**78.61**	**65.85**	**54.90**

5 对比实验

在本节中, 首先将本文方法与图像美学评价和情感分析的单任务基线方法进行比较. 同时, 分别将本文方法与现有的图像美学评价方法 DMA[15]和情感分析方法 DeepSentiBank[13]进行对比. 其次, 引入传统的基于参数硬共享的多任务学习算法作为对比方法, 该方法使不同任务共享浅层卷积层, 并在最后一个卷积层后进行分离, 以完成多个任务的预测. 此外, 还与近期提出的最先进的多任务卷积神经网络方法进行比较, 包括 cross stitch 网络[18]和 NDDR 网络[20]. 表 4 和表 5 展示了当不同方法选取 VGG16 和 ResNet50 作为基干网络时的实验结果. 在表 4 中, 本文方法在美学与情感任务的所有评价指标上都优于现有的单任务基线方法以及多任务学习方法, 超出单任务基线方法约 1%, 比两个任务的典型方法更优异. 同样, 在表 5 中, 本文方法实现了最佳性能. 这样的结果也说明相较于其他方法, 本文方法在目前常见的基干网络上均有一定程度的性能提升, 具有较强的稳定性.

表 4 基于 VGG16 网络的性能对比实验 (%)

方法	美学评价		情感分析	
	Accuracy	F_1	*Accuracy*	F_1
单任务基线	79.23	77.18	64.33	54.46
DMA	79.55	77.52	N/A	N/A
DeepSentiBank	N/A	N/A	64.51	54.54
参数硬共享网络	79.41	77.43	64.67	54.21
Cross stitch	79.82	77.48	65.24	54.65
NDDR	79.96	77.50	64.92	54.59
本文方法	**80.37**	**78.32**	**65.87**	**55.16**

表 5 基于 ResNet50 网络的性能对比实验 (%)

方法	美学评价		情感分析	
	Accuracy	F_1	*Accuracy*	F_1
单任务基线	79.28	77.31	64.36	54.27
DMA	79.43	77.49	N/A	N/A
DeepSentiBank	N/A	N/A	64.73	54.39
参数硬共享网络	79.39	77.38	64.42	54.09
Cross stitch	79.95	77.74	64.57	54.37
NDDR	79.86	77.52	64.87	54.81
本文方法	**80.55**	**78.61**	**65.85**	**54.90**

接下来, 本文讨论梯度重校准策略对美学与情感多任务学习过程中的作用及改善. 首先, 将本文方法与固定损失权重方法 (fixed weight) 进行比较. 固定损失权重方法需要通过网格搜索法来寻找合适的权重值, 并在实验前预先固定, 实际上很难准确的把握该权重. 因此在本次实验中, 本文固定各任务损失权重为 1. 此外, 本文将近期提出的经典的任务间梯度平衡策略作为对比方法, 包括 Uncertainty[36]、DWA[34]和 GradNorm[35]. 表 6 展示了不同方法的实验效果. 从表 6 可以看出, 固定损失权重方法在情感分析任务中与其他方法相比效果相差无几, 而在美学评价任务的两个评价指标上均表现最差. 原因在于在未加平衡的联合学习过程中, 情感分析任务的大量级梯度主导了多任务学习的反向传播过程, 使美学评价任务偏离了学习方向, 导致其性能下降. 其中 Uncertainty、DWA 和 GradNorm 均是通过考虑不同任务训练难度和收敛速度的差异性以确定各任务损失的权重, 且实验效果相差不大,

© 中国科学院软件研究所 http://www.jos.org.cn

可以得到一定程度的性能提升. 本文所提出的梯度重校准策略对交互模块回传的梯度进行重校准, 使其能与各网络分支进行安全整合, 可以更准确地使各任务梯度得到平衡, 并在对比实验中得到了最优性能.

表 6 梯度平衡策略对比实验 (%)

任务间平衡方法	美学评价		情感分析	
	Accuracy	F_1	*Accuracy*	F_1
Fixed weight=1	80.55	78.61	65.85	54.90
Uncertainty	80.72	78.78	66.21	55.19
DWA	80.62	78.95	65.64	54.95
GradNorm	80.75	78.71	65.90	54.87
梯度重校准	**81.15**	**79.09**	**66.34**	**55.62**

6 模型可视化

为了直观地理解本文所提出的自适应特征交互模块 (AFI 模块) 的效果, 本文应用 Grad-Cam 技术[43]可视化已训练好的模型. Grad-Cam 使用反向传播的梯度值来产生类激活图, 突出显示输入图像中用于预测目标类的重要区域. 通过观察重要区域, 本文可以深入了解模型是如何利用特定于任务的信息. 具体来说, 本文在模型最后一个 AFI 模块应用 Grad-Cam 方法, 分别生成美学和情感任务的可视化结果, 并与单任务基线方法的可视化结果进行比较. 图 5 展示了若干测试图像的可视化结果, 每个测试图像标注了真实的美学和情感标签, 以及目标类的 Softmax 预测分数 *P*. 图中美学评价和情感分析类激活图, 由自适应特征交互网络的最后一个 AFI 模块和单任务基线的最后一个卷积层生成.

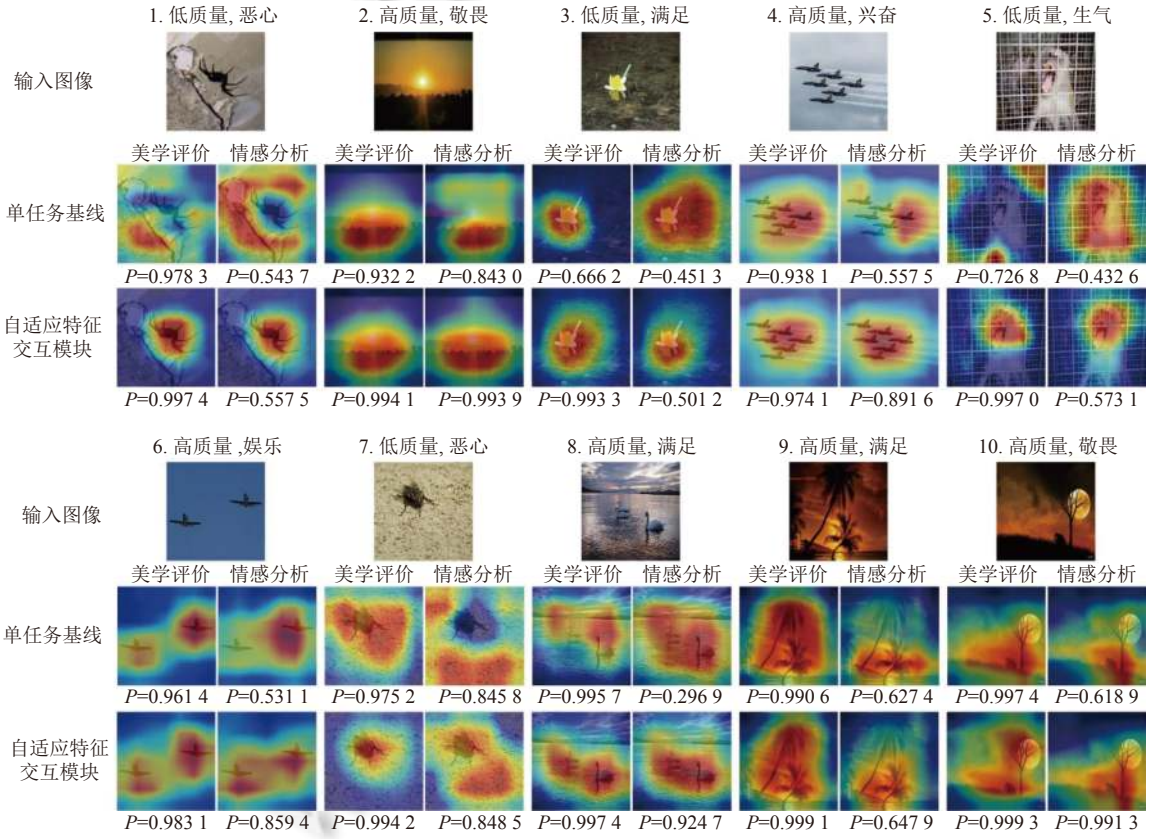

图 5 若干测试图像的可视化结果

© 中国科学院软件研究所 http://www.jos.org.cn

通过对比分析, 本文注意到 AFI 模块的一个明显优势是它能够聚焦重要的特征信息. 以图 5 中 2、3、4、8 这 4 幅图像为例, AFI 模块中测试图像的重要区域中心与单任务基线的重要区域所在位置差别不大, 但 AFI 模块能更聚焦重要区域而不分散. AFI 模块的另一个优势是促进了任务之间的知识转移, 即根据其他任务的特征来细化自身任务的特征. 以图 5 中第 5 幅图像为例, 美学评价单任务基线的重要区域并不包含图中动物的面部内容, 而关注于周边背景. 在经过 AFI 模块后, 情感分析任务的特征有效的转移到美学评价任务中, 使其覆盖于面部内容. 同样对于第 7 幅图像, AFI 模块将美学评价中的重要区域转移到情感分析中, 从而使更准确的区域得以高亮显示. 另外, 需要注意的是, 在实际的预测过程中, AFI 模块预测目标类的 Softmax 分数与单任务基线方法相比也实现了显著提高. 特别是对于第 2、4、5、6、8、10 等多幅图像, AFI 模块在情感分析任务中较大地提高了 Softmax 分数, 在 3、5 两幅图像中美学评价的 Softmax 分数也提升较高. 通过以上的分析, 本文可以发现通过 AFI 模块的自适应特征交互操作, 可以很好地细化特定于任务的特征, 并显著提升目标类的 Softmax 分数, 使每个任务的性能都实现较大的提升.

7 总 结

在本文中, 我们探索了图像美学与情感任务的内在联系, 通过引入含有自适应特征交互模块的深度多任务学习框架, 解决了统一的图像美学和情感预测问题. 在实验中, 本文注意到保持各任务的梯度平衡是至关重要的, 因此本文提出了一种新颖的任务间梯度平衡策略, 通过对多任务结构中任务间交互模块回传的梯度值进行重校准, 以保证美学和情感任务在统一学习的框架下平衡学习. 实验结果不仅验证了自适应特征交互模块对美学评价和情感分析任务的可行性, 同时也证明了本文所提出的梯度重校准策略的优越性. 此外, 本文创建了美学情感联合数据集 UAE, 首次将图像的美学和情感标签相关联, 可将其作为未来研究的基础. 在将来的工作中, 本文会进一步探索图像美学评价与情感分析任务间的深层联系, 并为其提供简单有效的任务间梯度平衡策略.

References:

[1] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[2] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

[3] Cui CR, Shen JL, Nie LQ, Hong RC, Ma J. Augmented collaborative filtering for sparseness reduction in personalized POI recommendation. ACM Trans. on Intelligent Systems and Technology, 2017, 8(5): 71. [doi: 10.1145/3086635]

[4] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[5] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

[6] Deng YB, Loy CC, Tang XO. Image aesthetic assessment: An experimental survey. IEEE Signal Processing Magazine, 2017, 34(4): 80–106. [doi: 10.1109/MSP.2017.2696576]

[7] Joshi D, Datta R, Fedorovskaya E, Luong QT, Wang JZ, Li J, Luo JB. Aesthetics and emotions in images. IEEE Signal Processing Magazine, 2011, 28(5): 94–115. [doi: 10.1109/MSP.2011.941851]

[8] Cui CR, Fang HD, Deng X, Nie XS, Dai HS, Yin YL. Distribution-oriented aesthetics assessment for image search. In: Proc. of the 40th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tokyo: Association for Computing Machinery, 2017. 1013–1016. [doi: 10.1145/3077136.3080704]

[9] Ren J, Shen XH, Lin Z, Mech R, Foran DJ. Personalized image aesthetics. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 638–647. [doi: 10.1109/ICCV.2017.76]

[10] Zhao SC, Zhao X, Ding GG, Keutzer K. EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: Association for Computing Machinery, 2018. 1319–1327. [doi: 10.1145/3240508.3240591]

[11] Kao YY, He R, Huang KQ. Deep aesthetic quality assessment with semantic information. IEEE Trans. on Image Processing, 2017, 26(3):

© 中国科学院软件研究所 http://www.jos.org.cn

1482–1495. [doi: 10.1109/TIP.2017.2651399]

[12] Cui CR, Liu HH, Lian T, Nie LQ, Zhu L, Yin YL. Distribution-oriented aesthetics assessment with semantic-aware hybrid network. IEEE Trans. on Multimedia, 2019, 21(5): 1209–1220. [doi: 10.1109/TMM.2018.2875357]

[13] Chen T, Borth D, Darrell T, Chang SF. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. arXiv: 1410.8586, 2014.

[14] Xu BH, Fu YW, Jiang YG, Li BY, Sigal L. Video emotion recognition with transferred deep feature encodings. In: Proc. of the 2016 ACM on Int'l Conf. on Multimedia Retrieval. New York: Association for Computing Machinery, 2016. 15–22. [doi: 10.1145/2911996.2912006]

[15] Lu X, Lin Z, Shen XH, Mech R, Wang JZ. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 990–998. [doi: 10.1109/ICCV.2015.119]

[16] Leder H, Belke B, Oeberst A, Augustin D. A model of aesthetic appreciation and aesthetic judgments. British Journal of Psychology, 2004, 95(4): 489–508. [doi: 10.1348/0007126042369811]

[17] Ruder S. An overview of multi-task learning in deep neural networks. arXiv:1706.05098, 2017.

[18] Misra I, Shrivastava A, Gupta A, Hebert M. Cross-stitch networks for multi-task learning. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3994–4003. [doi: 10.1109/CVPR.2016.433]

[19] Ruder S, Bingel J, Augenstein I, Søgaard A. Latent multi-task architecture learning. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 4822–4829. [doi: 10.1609/aaai.v33i01.33014822]

[20] Gao Y, Ma JY, Zhao MB, Liu W, Yuille AL. NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3200–3209. [doi: 10.1109/CVPR.2019.00332]

[21] Yim J, Jung, H, Yoo B, Choi C, Park D, Kim J. Rotating your face using multi-task deep neural network. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 676–684. [doi: 10.1109/CVPR.2015.7298667]

[22] Zhang ZP, Luo P, Loy CC, Tang XO. Facial landmark detection by deep multi-task learning. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 94–108. [doi: 10.1007/978-3-319-10599-4_7]

[23] Datta R, Joshi D, Li J, Wang JZ. Studying aesthetics in photographic images using a computational approach. In: Proc. of the 9th European Conf. on Computer Vision. Graz: Springer, 2006. 288–301. [doi: 10.1007/11744078_23]

[24] Dhar S, Ordonez V, Berg TL. High level describable attributes for predicting aesthetics and interestingness. In: Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011. 1657–1664. [doi: 10.1109/CVPR.2011.5995467]

[25] Tang XO, Luo W, Wang XG. Content-based photo quality assessment. IEEE Trans. on Multimedia, 2013, 15(8): 1930–1943. [doi: 10.1109/TMM.2013.2269899]

[26] Marchesotti L, Perronnin F, Larlus D, Csurka G. Assessing the aesthetic quality of photographs using generic image descriptors. In: Proc. of the 2011 IEEE Int'l Conf. on Computer Vision. Barcelona: IEEE, 2011. 1784–1791. [doi: 10.1109/ICCV.2011.6126444]

[27] Fang HD, Cui CR, Deng X, Nie XS, Jian MW, Yin YL. Image aesthetic distribution prediction with fully convolutional network. In: Proc. of the 24th Int'l Conf. on Multimedia Modeling. Bangkok: Springer, 2018. 267–278. [doi: 10.1007/978-3-319-73603-7_22]

[28] Borth D, Ji RR, Chen T, Breuel T, Chang SF. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proc. of the 21st ACM Int'l Conf. on Multimedia. Barcelona: Association for Computing Machinery, 2013. 223–232. [doi: 10.1145/2502081.2502282]

[29] Wang WN, Yu YL. A survey of image emotional semantic research. Journal of Circuits and Systems, 2003, 8(5): 101–109 (in Chinese with English abstract). [doi: 10.3969/j.issn.1007-0249.2003.05.023]

[30] Zhao JJ. Studies on related technologies of the mapping between visual features of images and emotional semantics [Ph.D. Thesis]. Taiyuan: Taiyuan University of Technology, 2010 (in Chinese with English abstract). [doi: 10.7666/d.d083132]

[31] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[32] You QZ, Luo JB, Jin HL, Yang JC. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. Austin: AAAI Press, 2015. 381–388. [doi: 10.1609/aaai.v29i1.9179]

[33] Kawakami R, Yoshihashi R, Fukuda S, You SD, Iida M, Naemura T. Cross-connected networks for multi-task learning of detection and segmentation. In: Proc. of the 2019 IEEE Int'l Conf. on Image Processing. Taipei: IEEE, 2019. 3636–3640. [doi: 10.1109/ICIP.2019.8803687]

[34] Liu SK, Johns E, Davison AJ. End-to-end multi-task learning with attention. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1871–1880. [doi: 10.1109/CVPR.2019.00197]

© 中国科学院软件研究所 http://www.jos.org.cn

[35] Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmässan: PMLR, 2018. 794–803.

[36] Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7482–7491. [doi: 10.1109/CVPR.2018.00781]

[37] You QZ, Luo JB, Jin HL, Yang JC. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Phoenix: AAAI Press, 2016. 308–314.

[38] Li X, Wang WH, Hu XL, Yang J. Selective kernel networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 510–519. [doi: 10.1109/CVPR.2019.00060]

[39] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

[40] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

[41] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[42] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics. Sardinia: JMLR, 2010. 249–256.

[43] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

附中文参考文献:

[29] 王伟凝, 余英林. 图像的情感语义研究进展. 电路与系统学报, 2003, 8(5): 101–109. [doi: 10.3969/j.issn.1007-0249.2003.05.023]

[30] 赵涓涓. 图像视觉特征与情感语义映射的相关技术研究 [博士学位论文]. 太原: 太原理工大学, 2010. [doi: 10.7666/d.d083132]

申朕(1995－), 男, 硕士, 主要研究领域为机器学习, 计算机视觉, 多任务学习.

余俊(1996－), 男, 博士生, 主要研究领域为机器学习, 计算机视觉, 美学质量评价, 多任务学习.

崔超然(1987－), 男, 博士, 教授, CCF 专业会员, 主要研究领域为信息检索, 推荐系统, 多媒体, 机器学习.

黄瑾(1994－), 女, 博士生, 主要研究领域为机器学习, 计算机视觉, 多模态融合.

董桂鑫(1997－), 男, 硕士, 主要研究领域为机器学习, 计算机视觉, 金融数据分析.

尹义龙(1972－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器学习, 数据挖掘, 模式识别, 生物特征识别.

© 中国科学院软件研究所 http://www.jos.org.cn