基于自动挖掘情感区域的视觉情感预测 (中文翻译版)*

杨巨峰 [1]　折栋宇 [1]　孙明 [1]　程明明 [1]　Paul L. Rosin [2]　王亮 [3]
[1] 南开大学　[2] 卡迪夫大学　[3] 中科院自动化所

摘要

目前，越来越多的用户通过社交网络上的图片和视频等媒介来表达自己的观点，从视觉内容上对情感进行自动评估已经开始受到越来越多的关注，本文旨在研究在识别过程中涉及高度抽象概念的视觉情感分析问题。现有的大多数方法都只关注于提升从图像全局角度捕获特征的表达能力，但通过观察可以发现图像的整体和局部区域都可以传达出重要的情感信息，受到该启本文提出了一个框架来有效利用图片中表达情感的区域，我们首先利用现有的目标检测工具去生成候选区域，再使用候选区筛选算法去除冗余和具有干扰的目标区域，然后利用卷积神经网络计算每一个候选区域的情感得分，通过同时考虑物体得分和情感得分自动发掘表达情感的区域。最后，将整张图片和局部区域的卷积神经网络输出相结合，产生最后的预测结果。由于标注情感区域非常主观并且费力，我们的框架只需要图像级别的标签，可以显著地减少在训练中需要的标注负担，这对于情感分析尤其重要。大量的实验表明我们提出的算法在八个主流的基准数据集上的结果均超过了目前最先进的方法。

关键词：情感区域　卷积神经网络
情感分类　视觉情感分析

1. 引言

随着社交网络的不断普及，越来越多的网络用户倾向于用不同的媒介去表达他们的观点 [64]，识别其中所蕴含情感的算法对于理解这些用户行为很有帮助 [35]，尤其是理解图片视频等视觉媒体内容中的情感已经在

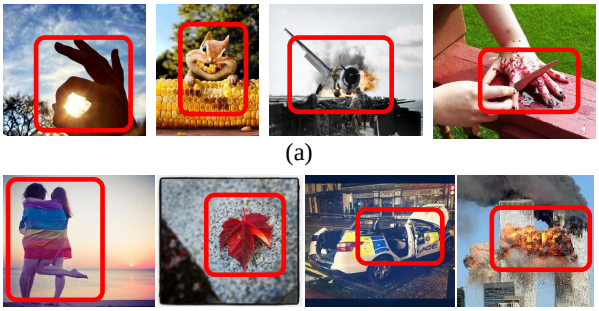

图 1. 来源于主流数据集的图片：(a) Twitter I [57] 和 (b) Twitter II [3]。边界框表示被用户标注的局部情感区域，由此看出，情绪由情感区域和整张图片的表现共同引发。

研究领域引起了越来越多的注意，该种分析算法的潜在用途是非常广泛的，主要包括情感图片检索 [36]、美学质量分类 [30]、意见挖掘 [59]、评论助手 [9]等应用。

受到心理学和艺术原则的启发，目前的研究工作已经调查了不同的手工图像特征对于情感分类的影响（如颜色 [43, 65]，纹理 [55, 32]，形状 [31]等），旨在赋予计算机和人类一样感知情感的能力。卷积神经网络 [71]可以自动地学习图片的深度特征，而不需要手工地设计视觉特征，一些研究者已经将卷积神经网络应用到了图片情感分类的问题中 [5, 57, 4, 58]，并说明了其相对于人工设计的特征在情感分类任务上表现更好。

视觉情感分析从本质上而言比传统的识别任务更有挑战性，因为它涉及更高层次的抽象，以及人类识别过程中的主观性 [23]。相对于物体分类 [20]，场景识别 [72]等任务而言，从社交媒体上识别图片引发的情感比其他很多视觉识别任务更难，因而对于视觉情感问题而言，考虑一些丰富的因素是必要的。目前大多数现存的方法尝试用卷积神经网络从整个图像的全局角度学习情感表征，然而我们观察发现视觉情感往往也由

*本文工作代码已开源至https://github.com/sherleens/AR_discovery

图像的局部区域引发 [38, 56, 28]。不同于具体的视觉物体检测任务 [42]，情感建模在低层视觉特征和高层情感之间存在 “情感鸿沟” [32]。

目前很少的情感分析工作会关注局部信息的使用，Li 等人 [28]提出了一个基于两层稀疏表示并且同时考虑局部和全局信息的情境感知分类模型。然而这种方法主要受限于其严重依赖于依据物体外观进行的初始分割结果。除此之外，他们还假设所有的区域对于情感预测都具有相同的权重，这与人类视觉系统会有选择地处理图像部分细节这一关于人类注意力的理论相矛盾 [12]。You 等人 [56]尝试将图片的局部区域和描述性视觉属性匹配，旨在发现特定属性的区域，但是对于情感分析来说缺少泛化能力。

为了解决这个问题，我们提出了利用局部细节和全局信息来进行视觉情感分析。我们引入一个新的概念名为情感区域（AR, Affective Region），包括两个鉴别特征：

1. AR 是一个显著性区域，可能包含一个或更多个吸引人注意力的物体；
2. AR 可以传达出重要的情感；

图 1 展示了在广泛使用的数据集 [57, 3]上的一些 ARs，可以观察到视觉情感可以由图片当中的情感区域所引发。例如，（a）中的第四张图片中的情感主要来自于流血的手这部分区域；（b）中第二张图片中鲜艳的叶子传递了积极的情绪，而与灰色石头无关。然而手动地对图片中的情感区域进行标记并训练相应的检测器不仅太主观而且很耗费劳动力。本文工作提出了一个框架只需要图片级标签的框架就能自动地发掘情感区域，从而有效地降低标注负担。

详细而言，由于物体和情感的强共存关系 [8]，对于输入图片我们首先用一个现成的工具去生成候选边界框和物体检测得分。然后利用一个候选区域筛选算法在保留一些有价值的区域的情况下去除冗余的目标区域。将每一个候选区域送入卷积神经网络从而计算情感得分。接下来结合物体得分和情感得分来计算 AR 得分，基于以上步骤同时考虑似物性和情感传达程度，通过重排候选区域，找到前 K 个情感区域作为最终的检测结果。最后，将来自全局视图和局部视图的卷积神经网络输出结果通过一些融合方式（如最大值池化、和值池化和串联连接的方式）来得到最后的预测类别。

综上，我们的贡献总结如下：

- 我们提出了一个深度框架去自动发现图片中很有可能诱发重要情感信息的情感区域。该框架不依赖于对象类别并且不需要边界框标注，比现有的方法更有普适性。
- 我们用一个深度卷积神经网络建立了一个视觉情感预测模型，同时利用了分别来自整张图片和局部区域的全局信息和局部信息。最后的表征对于视觉情感分类是有效的，并且在情感数据集上的表现超过了现存最先进的方法。
- 实验结果说明了我们提出的框架在迁移学习的帮助下可以推广到小规模的基准数据集上。

该期刊论文从四个方面扩展了我们更早时候的工作 [48]：（1）通过加入候选区选择模块去抑制可能出现的噪声区域并且降低计算负荷。（2）利用三种融合操作将全局表征和情感区域合并。（3）提供了更多的实验细节并且展示了在大规模数据集和小规模数据集上更充分的实验结果，并系统地探讨了其中的超参数的选择。（4）在情感基准数据集上评估发现的情感区域和类别标签的一致性，展示了我们提出的方法在没有人工标注的情况下可以自动地找到高质量的 ARs。

本文其余部分的安排如下所示：第 2 节总结了视觉情感分析和深度学习领域最近的工作。第 3 节介绍了为检测情感区域提出的方法和用来预测情感的深度框架。在第 4 节和第5节，我们展示并可视化了在主流数据集上的实验结果。最后，第 6 节总结了这篇论文。

2. 相关工作

目前已提出了很多基于图像 [32, 3]和视频 [2, 24]的情感分析方法，本部分主要回顾了情感图片预测方法以及和我们工作密切相关的基于区域的深度模型。

2.1. 情感图片分类方法

情感图片的预测方法大致可分为：基于维度的方法和基于类别的方法。基于维度的方法在二维坐标空间（即极性-唤起程度空间）[34]或者三维坐标空间 [47]上表示情感。Hanjalic 等人 [19]利用极性、唤起程度和可控程度三个基准维度来表示人类情感反应，在该三维空间中可以找到每一个情感状态的对应值。Zhao 等 [68, 66]将共享稀疏回归作为学习模型在二维空间中预测个人对

于图片的情感认知。同时，分类方法将情感映射到一些典型的类别中。还有一些工作预测情感类别的离散概率 [69, 53, 54, 63, 67]。由于基于类别的方法更为直观且容易让人理解，本文工作主要以情感类别预测作为目标。

2.1.1 浅层模型方法

大多数传统情感图片预测的方法尝试利用低层的特征来分类，如 Machajdik 等 [32]基于艺术和心理学理论定义了一个含有丰富的手工特征的组合，包括成分、颜色变化和图片纹理等。Lu 等 [31]研究了自然图片中的形状特征怎样影响人类情感的产生，并实验证明了圆形-棱角和简易-复杂对预测情感内容的不同的作用。Zhao 等 [65]引入了根据艺术原则而设计的更鲁棒和更稳定的视觉特征。并证明了这些手工视觉特征在一些从特定领域筛选的图片构成的小数据集上的分类效果较好，比如抽象画数据集和艺术照片数据集 [32]。

为了衔接低层特征和高级感情之间的“情感鸿沟”，Borth 等 [3]建立了一种中层概念——形容词名词词组（ANPs），可以用来检测图片中的概念从而间接地表达情感。Li 等 [29]进一步计算了描述图片的 ANPs 的文本情感值加权和，同时考虑了文本情感的概念。Yuan 等 [60]提出了 Sentribute，即一个基于 102 维中层属性的图片情感分析算法，更具有可解释性且利于从高层次进行图像理解。此外，Zhao 等 [70]组合了不同级别的特征，包括源于艺术元素的低层特征，来自艺术规则定义的中层特征和多图学习架构中语义概念检测器得到的高层特征。Chen 等 [8]建立了目标检测模型去识别六种常见的目标，包括车、狗、连衣裙、脸、花朵和食物，并且提出了一个新的分类模型去处理属性和语义概念之间的相似性。相反，我们的算法只关注筛选出来的区域是否包含物体，无关物体的类别，在实际应用中有更强的鲁棒性。

2.1.2 深度模型方法

近几年，卷积神经网络已经被运用于多种领域下的视觉识别系统中 [25, 16]，相比于更传统的识别流程还需要将计算图片中手工设计的特征作为预处理步骤，这些模型的强大之处在于可以利用反向传播算法 [27]从原始数据中学到有辨识度的特征。

数据集	矩形框标注	积极	消极	总计
IAPSa [33]	N	209	186	395
Abstract [32]	N	139	89	228
ArtPhoto [32]	N	378	428	806
Twitter I [57]	N	769	500	1,269
Twitter II [3]	N	463	133	596
EmotionROI [39]	Y	660	1,320	1,980
Flickr&Instagram [58]	N	16,430	6,878	23,308
Flickr [3]	N	435,798	48,424	484,222

表 1. 当前公开的情感数据集统计。可以看出，该领域的大部分数据集都只包含不到 2000 个样本，主要因为标注过程的主观性以及耗时费力。这里，Flickr 数据集是利用网络数据进行弱标记的数据，且除了 EmotionROI 数据集之外都没有提供对应于情感区域的矩形框标注信息。

最近一些方法利用卷积神经网络进行图片情感预测，比如 Chen 等人 [7]基于他们之前的工作 [3]，将深度网络应用于构建一个用于视觉情感概念的分类模型——DeepSentiBank，此模型在标注准确率和检索表现上都展现了显著的提升。此外，一些方法利用了在大规模普通数据集 [11]上学习到的模型权重，然后对于视觉情感预测任务 [5, 4]进一步微调网络。在 [5]中，将卷积神经网络中 fc7 层的 4096 维输出和 fc8 层的 1000 维输出作为两类图像级别的特征用于分类。You 等 [57]使用大约 50 万被网站元数据标注的图片，提出一种渐进革新的训练策略得到了一个卷积神经网络，并进一步在 Flickr and Instagram(FI) 数据集上完成了基准分析。 [56]中提出了一个基于注意模型的方法，该模型考虑了由情感相关的视觉属性引起的局部视觉区域。

由于手工标注情感标签非常昂贵，现有的情感数据集包括 IAPSa[33]，ArtPhoto[32]，Abstract Paintings[32]，Twitter [57]，Twitter [3]和 Emotion-ROI[39]基本都少于 2000 张图片（如图 1 所示），这和训练一个鲁棒性的深度模型所需要的数据规模相距甚远。其中，Flickr 是利用图片上传者提供的元数据进行两类标签弱标记的数据集。而且，只有 EmotionROI 数据集提供了情感区域的标签。在这篇论文中我们专注于两类情感（积极情感和消极情感）预测问题，在有可靠标签的很多基准数据集上都可以验证我们方法的有效性。

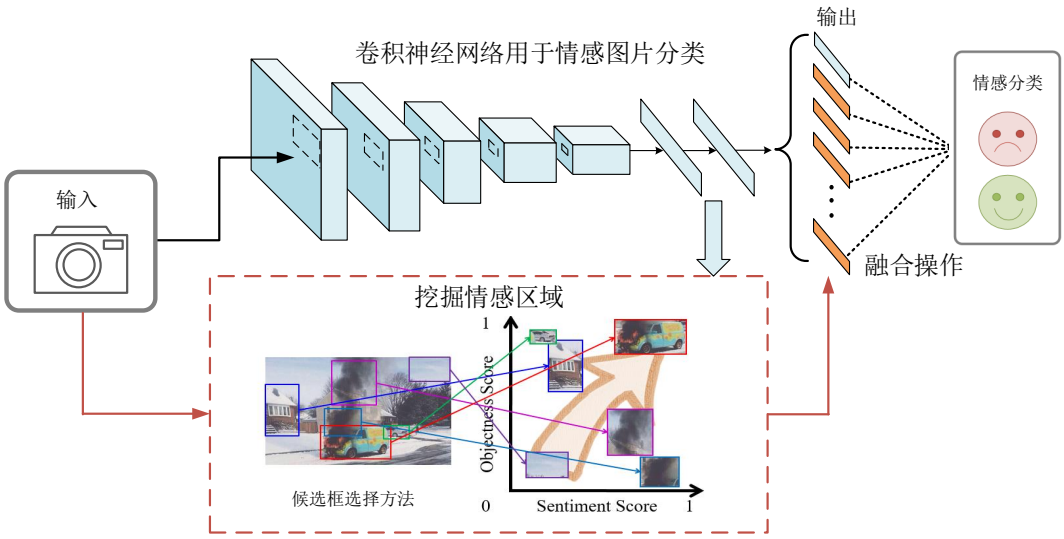

图 2. 本文提出方法的流程图，给定输入图片，我们可以得到数千个带有物体得分的候选区域，用候选区筛选算法去除一些重叠的或不重要的候选区。每个目标区域的情感得分可以通过卷积神经网络粗略地计算出来，然后结合物体得分去发掘情感区域。最后，通过使用几个可选择的操作将局部信息与整体表示融合，从而预测情绪标签。

2.2. 基于区域的深度方法

我们先回顾基于区域的深度 [16]方法的起源，R-CNN 是一种为了定位和分割物体利用卷积神经网络自下而上生成目标区域的算法。当被标记的训练数据不足，首先对于辅助任务进行有监督地预训练，然后在特定领域上进行微调，模型在最终任务上地表现就会显著提升，如 [17]中所证明的。Girshick[15]使用 Fast R-CNN 的方法证明了在提高检测准确率并且简化训练过程的情况下进一步削减训练和测试时间是可能的。Fast R-CNN 基于网络架构，去除了目标区域的计算过程，将检测时间降到了每张图片 50-300 毫秒。Ren 等 [42]引入了在每个位置上同时预测目标边界和目标检测评分的全卷积网络。同时，R-CNN 已经被应用到了很多任务，如行人检测 [62]、动作检测 [40, 18]和语义分割 [14]等。

与基于区域卷积神经网络在图片中寻找显著物体的传统方法不同，我们工作的目标是自动识别引发情感的 ARs 并且利用局部信息作为情感表示的补充。这不仅要求我们分析包含物体的区域而且要考虑对所选区域有情感方面影响的周围背景 [38]。而且，基于 R-CNN 的方法需要边界框标注信息进行训练，但手工标注情感区域是非常耗时费力的。在本文中，我们借助了现成的工具生成物体区域作为候选情感区域并且根据低层和情感层面的内容选择 AR。和需要精确分割 [28]或者具体类别信息 [8]的方法相比，我们这个只在预处理阶段需要目标区域的方法更加简单，并且能更好的推广到其他数据集。

3. 方法

在本节中，我们的目标是提出一个算法自动识别传达重要情感的 ARs 并且将标准的全局表示和局部表示结合进行图片情感分析。图 2 展示了框架的流程图，我们首先用目标检测框架 Edgeboxes [73]生成候选窗来引导发掘 ARs，然后利用候选选择方法降低冗余和噪声区域。在检测 ARs 的过程中，每个目标区域的情感内容在低层和情感层面上都会被估计。最后，被检测出来的 ARs 的深度表示通过三种可选择的融合方式（最大值池化、和值池化和串联连接）和全局表示进行组合产生最后的预测结果。

3.1. 产生候选区域

3.1.1 生成过程

检测具体的视觉目标（如狗和车）在计算机视觉领域已经得到了广泛的研究 [14, 41]。然而，对抽象的情

感概念，比如愉悦和兴奋，进行建模是非常有挑战性的，难点主要来源于低层视觉特征和高层情感之间“情感鸿沟”。之前的方法 [3, 8]已经证明了将形容词和具体的目标结合在一起组成的这种视觉概念对于视觉情感分析来说更容易检测和处理。受物体和情感之间这种强共现关系的启发，我们认为物体区域可以被看作潜在的情感区域。

因为我们的框架输入物体目标区域，通过把每个情感区域的预测结果和全局表示及进行融合得到最后的预测，所以该框架的表现在很大程度上依赖于候选区域的质量。然而，一个有效的候选区域的提取方法是很具挑战性的，因为情感区域检测不仅需要捕获物体而且还要考虑背景中可能引发情感的区域。有两个条件需要满足，首先，该架构基于候选目标区域能够覆盖情感图片中的物体和部分背景这一假设，因此需要一个较高的检测召回率。第二，由于选出的情感目标区域送入卷积神经网络，所以只能有有限数量的候选区域产生，以便在保持准确性的同时提高效率。

在过去的几十年，出现了很多物体区域的方法来处理目标检测问题。[22, 21]中说明 EdgeBoxes [73]和 BING [10]比起 Selective Search [49]和 Objectness [1]这些方法而言更快，而且和 BING 相比，EdgeBoxes 能实现更好质量的目标区域。考虑到速度和质量之间的平衡，本文用 EdgeBoxes 生成一系列的候选窗口考虑到更好的权衡。这样一个现成的工具可以在一秒内生成数千个候选框，随后基于对象边界估计的细化步骤被应用于提高定位效果。对于一个给定的图片 I，由 EdgeBoxes 可以产生一系列的带有物体得分 $B = \{b_i; Obj\_score_i^I\}_{i=1}^n$ 的候选边界框。

3.1.2 筛选和过滤

对于目标检测，为了获得高召回率，Zitnick 等 [73]用了一种自下而上的策略，在每张图片中生成数千个物体候选区域。然而，大多数的候选区域严重的重叠，对于预测情感来说是冗余的。过滤掉只携带少量情感的噪声目标区域是必要的，而且在算法的初级阶段去除噪声区域也可以减少后续步骤的计算时间。为了解决这个问题，我们受 [50]的启发，引入了候选区域筛选模块从情感候选区域选择目标区域。

对于每张在物体检测中可以实现高召回率的图片，EdgeBoxes 生成了几千个目标区域。然而，因为它仍然生成了需要卷积神经网络处理的大量原始的物体窗口 [52]，我们首先检查候选边界框相同的几何特征（面积和高宽比）。我们根据经验过滤掉小面积的区域 (< 800 像素) 或者高宽比 (或宽高比) 超过了临界值（> 6）的区域, 因为太小或者太长的物体不太可能引起人们的注意。因此，目标区域就会减少很多，然后用候选区筛选方法对它们进一步处理。依据之前的算法 [13, 50]，我们对每张图片建立了关系矩阵 $W \in \mathbb{R}^{n\times n}$。

在矩阵中，每个元素表示任意一对边界框之间的交集和并集之比（IoU），n 代表候选区域的数量：

$$W_{ij} = \frac{|b_i \cap b_j|}{|b_i \cup b_j|}, \tag{1}$$

其中 $|\cdot|$ 用来度量像素的数量。然后我们用归一化分割算法 [44] 将候选边界框分成 m 簇。详细而言，标准化的图拉普拉斯矩阵通过 $L = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$ 可以计算。其中 $D \in \mathbb{R}^{n\times n}$ 是一个对角矩阵，而且 $D_{ii} = \sum_{j=1}^{n} W_{ij}$。特征向量矩阵 $V = [v_1, \cdots, v_m] \in \mathbb{R}^{n\times m}$ 由 L 的 m 个最小的特征向量 $\{v_1, \cdots, v_m\}$ 构成。最后利用 k-means 聚类算法获得 m 个类别标签，V 中的每一行是对应样本的特征 [44]。如图 3 所示，边界框首先要通过过滤减少计算量。然后用 m 簇的边界框，我们在每一簇中挑选最高物体检测得分的区域作为目标区域并且为每个图片生成 m 个候选区域 $H = \{h_i\}_{i=1}^m$。和广泛用于过滤的贪婪非最大值抑制 (NMS) 方法 [16]相比，我们的候选区域筛选方法在去除冗余和噪声边界框的时候可以生成特定数量的目标区域。

3.2. 挖掘情感区域

3.2.1 初始化框架

卷积神经网络在相关的计算机视觉任务上实现了目前最好的表现，比如通过微调在 ImageNet 数据集上预训练好的模型进行美感质量评价 [30]、图片风格识别等。在本文工作中，我们使用的卷积神经网络是基于 16 层的深度模型 VGGNet [46]。为了将在 ImageNet 数据集上预训练好的模型适应于情感分析的任务上，卷积神经网络首先在目标情感数据集（比如 Flick and Instagram）上微调，利用原始图片（没有任何边界框）去调整深度模型的参数。作为一个监督学习的方法，目标是用微调过的卷积神经网络模型从情感样本训练集 $\{(I_i, l_i)\}_{i=1}^N$ 中学习一个函数：$f : \mathcal{I} \to \mathcal{L}$, 其中 N 是训

图 3. 给定输入图片 (a)，由 EdgeBoxes 生成候选窗口，宽高比太小或太高的都会被过滤掉，如 (b) 所示，蓝色边界框内的目标区域在该步骤中被过滤掉的。(c) 中不同的颜色表明通过正则化分割产生的不同的簇，从中会选择有代表性的目标区域。

练集的大小，I_i 是相关的情感标签。在标准的训练过程中，通过优化传统的分类损失来最大化正确类别的概率 [25, 46]。令 d_i 是倒数第二层的输出，然后通过最小化如下所示的 softmax 损失函数微调最后一层:

$$l(\mathrm{W}) = \sum_{i=1}^{N}\sum_{j\epsilon l} 1(l_i = j)\log p(l_i = j|\mathrm{d}_i, \mathrm{w}_j), \qquad (2)$$

其中 $\mathrm{W} = \{\mathrm{w}_j\}_{j\epsilon l}$ 是模型参数的集合，如果 s 为真，指示函数 $1(s) = 1$，否则，$1(s) = 0$。每一个情感标签 $p(l_i = j|\mathrm{d}_i, \mathrm{w}_j)$ 可以通过 softmax 函数被定义为:

$$p(l_i = j|\mathrm{d}_i, \mathrm{w}_j) = \frac{\exp(\mathrm{w}_j^T \mathrm{d}_i)}{\sum_{j'\epsilon l}\exp(\mathrm{w}_{j'}^T \mathrm{d}_i)} \qquad (3)$$

3.2.2 估算情感得分

为了从情感层面评测候选目标区域的质量，我们将目标区域送入卷积神经网络来计算情感得分。对于输入图片 I 生成的情感候选区域 $H = \{h_i\}_{i=1}^m$，令最后一层的输出向量 $\{y_{ij}\}_{j=1}^c$ 表示第 i 个目标区域蕴含第 j 类情感的概率，c 设置为情感的类别数，本文中为 2。如果对于每一种情感的预测值是相近的，那么说明区分目标区域引发的情感是比较困难的。因此，我们的目标是只保留那些含有明显情感倾向的目标区域。我们定义了概率抽样函数从情感层面的角度去评估第 i 个区域的情感得分:

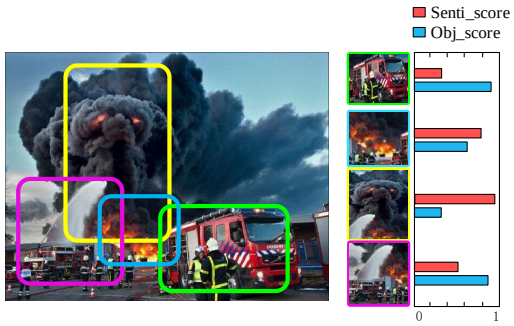

图 4. 一张图片中物体分数和情感分数的可视化。比如，具有高物体得分的区域说明相对应的边界框中极有可能是一个物体。

$$Senti\_score_i^I = \sum_{j=1}^{c} y_{ij} * \log y_{ij} + 1, \qquad (4)$$

对于二分类，得分范围从 0 到 1。 式 4 中定义的信息熵表示了预测情感的不确定程度，这和对情感区域进行从情感层面进行估计是一致的。和传统的方法相比，$Senti\_score_i^I$ 是高层更具语义性的度量。

3.2.3 筛选表达情感的区域

我们依据两个方面选择情感区域: i) 该区域包含物体的可能性，表示为 $Obj\_score_i^I$; ii) 在高级语义层面上，区域包含了多少情感，表示为 $Senti\_score_i^I$。图 4 说明了一个情感区域应该同时有高的 $Obj\_score_i^I$ 和 $Senti\_score_i^I$。因为 $Obj\_score_i^I$ 只能基于纹理特征度量区域包含物体的概率，但缺少情感信息方面的引导。$Senti\_score_i^I$ 在情感层面反应了图片的感情，而且能够移除对情感分析基本没有影响的大量噪声区域。这个分数赋予了物体区域一定的灵活性，当然也可以是在背景区域。考虑到每个得分的特点，我们引入了 AR_score 去估计每个区域的情感质量，定义如下:

$$AR\_score_i^I = \sqrt{(1-\alpha) * Obj\_score_i^{I\,2} + \alpha * Senti\_score_i^{I\,2}}, \qquad (5)$$

其中，α 在低层和情感层面之间进行权衡。在本文中，我们通过在大规模情感数据集上交叉验证选择 α。具

Algorithm 1 用情感区域进行视觉情感分析

Input:

输入图片: I

所需情感区域的数量: K

Output:

预测情感标签: $\vec{Y}$

1: 生成 n 个边界框和它们对应的物体得分 $B = \{b_i; Obj\_score_i^I\}_{i=1}^n$。
2: 用筛选候选区域的方法生成 m 个目标区域 $H = \{h_i\}_{i=1}^m$.
3: 用预先训练的卷积神经网络来初始化框架。
4: 令 $\vec{Y}_{Global}$ 表示对整张图片的预测。
5: 通过卷积神经网络模型将 H 从第二层传到最后一层。
6: 令 $y \in \mathbb{R}^{m \times c}$ 表示卷积神经网络模型中 m 个目标区域的情感概率, 计算式 4 中的情感得分。
7: 用式 5 计算每一个区域的 AR 得分。
8: 根据 AR 得分排序，选择前 K 个作为情感区域。
9: 用跨候选集的池化操作得到预测标签 $\vec{Y}$ 。
10: return $\vec{Y}$

有高 AR_score 的目标区域将被认为是 AR 并且用于情感预测，同时将 AR_score 低的目标区域从候选集合中去除。

3.3. 情感分类

基于初始化的架构，给定图片的情感分类可以总结如下。给定一个测试图片，我们首先基于 EdgeBoxes 生成情感候选区域。为了减少冗余，我们利用了基于 IoU 得分的筛选候选区域方法只保留最好的候选区域。同时基于物体检测得分和情感得分的选择能够吸引人注意力并且包含情感内容的情感区域。然后，对于每个目标区域和整个图片，通过卷积神经网络获得一个 c 维的预测结果，随后进行融合得到最后的预测。特别地，我们考虑了三种策略，分别为最大值池化、和值池化和串联连接。我们利用了跨候选集的池化操作将卷积神经网络的输出融合到综合预测。利用最大值池化，那些有较高预测得分的情感候选区域将会被保留，同时噪声区域被忽略。给定图片的情感概率 $\vec{Y}$ 可以按照如下定义：

$$\vec{Y} = \max\left(\vec{Y}_{Global}, \{\vec{Y}_{AR_j}\}_{j=1}^K\right), \tag{6}$$

其中，$\vec{Y}_{Global}$ 代表整张图片的预测，$\vec{Y}_{AR_j}$ 表示第个情感区域的预测，我们基于式 5 选择前 K 个情感区域，$\vec{Y}$, $\vec{Y}_{Global}$ 和 $\vec{Y}_{AR_j}$ 共享相同的向量结构 (y_{pos}, y_{neg}),y_{pos} 和 y_{neg} 分别表示积极情感和消极情感的预测概率。

和值池化融合了所有目标区域的预测概率，在这些区域中，一致推荐的权重可以被加强。

$$\vec{Y} = (1-\beta) * \vec{Y}_{Global} + \beta * \frac{1}{K} * \sum_{j=1}^{K} \vec{Y}_{AR_j}, \tag{7}$$

其中，β 在全局和局部预测之间进行权衡。而且 β 是在大规模情感数据集上进行交叉验证估算得到的。最大值池化和和值池化都可以生成情感概率作为最后的预测。

串联连接是通过合并特征形成综合表示，是一种简单并且有效的方法：

$$\vec{Y} = \left[\vec{Y}_{Global}, \{\vec{Y}_{AR_j}\}_{j=1}^K\right] \tag{8}$$

最后的特征由串联所有的预测结果生成，$\vec{Y}$ 的维度是 $(K+1) \times c$。在我们的实验中，我们设置所有样本的情感区域数量都是相同的，因此可以利用 SVM 算法对串联得到的特征进行分类。

4. 实验结果

本节呈现了我们的实验，将提出的方法在和最好的深度方法进行评估，验证了我们的框架在情感分类和情感检测任务中的有效性。

4.1. 数据集

我们在八个广泛使用的数据集上评估了我们提出的方法，数据集包括 IAPSa [33]，ArtPhoto [32]，Abstract Paintings [32]，Twitter I [57]，Twitter [3]，EmotionROI [38]，Flickr [3]和 Flickr and Instagram(FI) [58]数据集。我们根据图片的数量将数据集分为大规模数据集和小规模数据集，如图 5 所示。

4.1.1 小规模数据集

国际情感图片系统 (IAPS) [26]是一种在视觉情感分析研究中 [32, 65, 51, 55]广泛使用的数据集。IAPSa

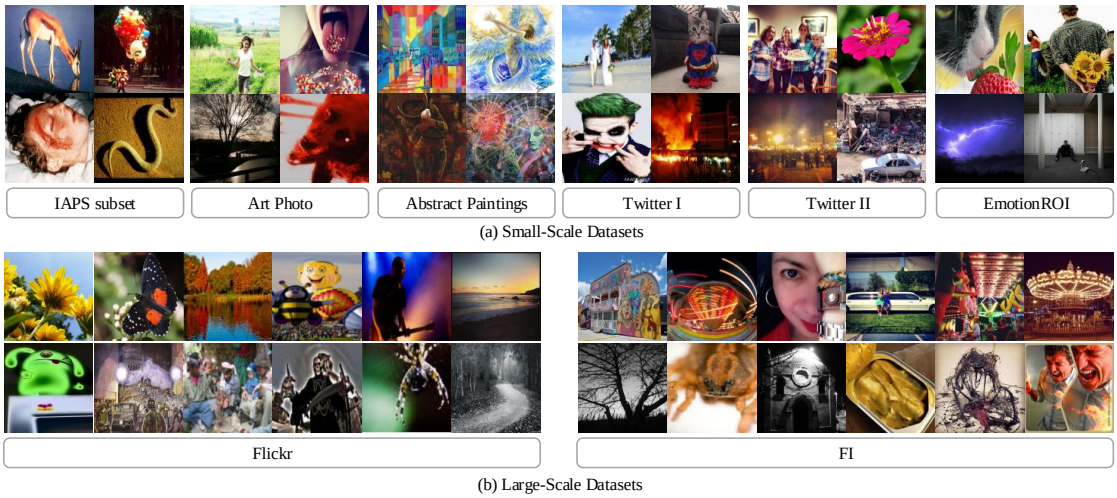

图 5. 来自 (a) 小规模和 (b) 大规模情感数据集的图例，这些图片来自于很多的领域包括艺术、现实生活、抽象画等，它们的数据分布都是不同的。

从 IAPS 中选择 395 张图片，然后利用 Mikels 的 8 类情感类别标注。ArtPhoto 包括来自照片分享网站上的 806 张艺术照片，被标注的标签由每张图片的主人提供。Abstract Painting 包含 228 个由颜色和纹理组成的对等抽象绘画。Twitter I 是从社交网站上搜集得到的共 1269 张图片，然后由亚马逊土耳其机器人平台（AMT）标注成两类（积极情感、消极情感）。我们在 Twitter I 的所有三个子集上以和 [57]相似的方式测试了我们的方法，三个子集分别是 “Five agree”, “At least four agree”, “At least three agree”, “Five agree” 表示对于给出的图片所有的五个 AMT 工作人员都给出了相同的情感标签。Twitter II 包括来自 Twitter 网站上的 603 张图片，标签也是通过 AMT 进行手工标注，最后的得到 470 个积极情感标签和 133 个消极情感标签。EmotionROI 是从 Flickr 上搜集，产生了有六个情感类别、1980 张图片构成的情感预测基准数据集。他们使用 AMT 来收集 15 个对区域的反应，这些反应唤起了情感，并代表了正确标签，假设每个像素对唤起情感的影响与覆盖那个像素的矩形的数量成比例。

4.1.2 大规模数据集

FI 是目前最大的已标注数据集，是通过用 8 个情感类别作为关键词从社交网站上搜索并收集产生的。共 225 个 AMT 工作人员进行标注数据，最终 23308 张图片得到了至少三个员工的一致标注。我们将 FI 分成了和 IAPSa 一样的两类别的数据集。Flickr 总共包括 484258 张图片，每张图片用相应的 ANP 进行自动标注。

因为关注的是两类情感预测，所以我们根据数据集的唤起程度将除了 Twitter I、Twitter II 和 Flickr 本来就被标注成两类情感的数据集之外的多情感标注转换成积极和消极两类情感。明确地讲，对于 IAPSa、ArtPhoto、Abstract Paintings 和 FI，我们根据 [33]将 Mikel 的八类情感类别分成两类，也就是说 amusement, awe, contentment 和 excitement 属于积极情感，anger, disgust, fear 和 sadness 属于消极情感。EmotionROI 根据 VA 得分被标注为七类情感，即 anger, disgust, fear, joy, sadness, surprise, neutral，其中 anger,disgust,fear,sadness 可以近似的被看作是消极情感。因为 joy 和 surprise 图片集合的平均 valence 比消极类别图片的平均 valence 高，我们将它们看作积极情感。在实验中我们不考虑中立情感的图片。

4.2. 实现细节

卷积神经网络有能力结合在更综合的数据集上学习的模型权重，这对于缺乏充分训练数据的任务是非常实用的性质。我们用 16 层 VGGNet [46]作为我们的基础架构。根据之前的工作 [5]，用在 ImageNet 上训

对比方法		FI	Flickr
基准方法	AlexNet [25]	60.54	55.13
	VGGNet [46]	70.64	61.28
	Fine-tuned AlexNet	72.43	61.85
	Fine-tuned VGGNet	83.05	70.12
	PCNN (VGGNet) [57]	75.34	70.48
	DeepSentiBank [7]	61.54	57.83
所提方法	obj + concatenation	83.85	70.05
	senti + concatenation	84.07	70.10
	AR + concatenation	84.83	70.51
	AR + sum-pooling	84.50	70.46
	AR + max-pooling	84.21	70.49
	AR + concatenation ($K = 8$)	86.35	71.13

表 2. 大规模数据集的测试集分类准确率，分别是 FI 和 Flickr。我们将自己提出的方法和不同的深度模型包括 ImageNet 模型（1-2 行），微调过的模型（3-4 行）和目前最先进的方法（5-6 行）进行了比较。我们的方法在不同配置下也给出了结果，包括与 top-1 的区域结合 (7-11 行) 以及利用更多的情感区域 (第 12 行)。obj/senti 说明只使用了物体得分/情感得分，而 “AR” 方法同时考虑物体得分和情感得分进行选择情感区域。

练好的权重初始化我们的模型。然后将 fc8 分类层从 1000 输出通道改为 2 输出通道，用大规模数据集微调预训练的网络，数据被随机划分，其中 80% 用于训练，5% 用于验证，15% 用于测试。卷积层和最后一层全连接层的学习率分别被初始化为 0.001 和 0.01。我们通过整个网络的随机梯度下降来对所有层进行微调，批量大小设置为 64。为了提取更精确的情感相关的信息，共进行 100000 次迭代来更新参数。我们所有的实验都是基于 NVIDIA GTX 1080 GPU 实现。对于候选区域筛选方法，考虑到性能和计算时间之间的权衡，我们对于每张图片设置 $m = 50$，对于情感区域提供了初始的候选区域。

在迁移学习的帮助下，我们还将该架构应用到了训练样本很有限的小规模数据集上。详细而言，我们将在 FI 数据集上训练好的卷积神经网络参数用到其他数据集上，在其他数据集的训练集上微调这个模型。除了那些特定的训练集/测试集划分 [3, 39]，我们将小数据集随机划分，80% 的数据用于训练，20% 的数据用于测试，然后我们用 5 折交叉验证进行实验取平均准确率作为最后的结果。

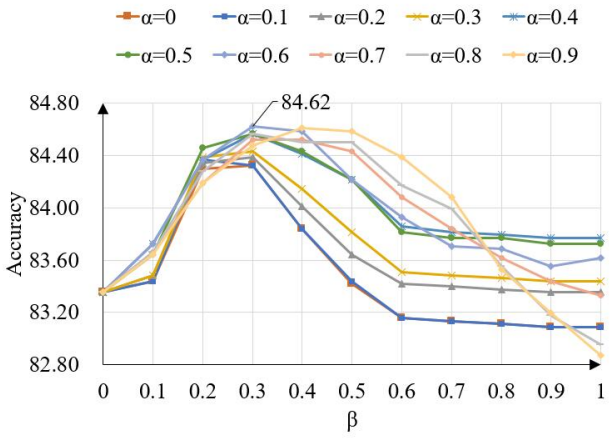

图 6. 不同的 α 值和 β 值对于 FI 验证集的影响。我们在接下来的实验中令 $\alpha = 0.6, \beta = 0.3$ 。

4.3. 基准方法

在下面这个小节中，我们将本文提出的方法和对于图片情感预测最先进的算法进行评估，包括那些基于手工特征和深度模型的方法。除此之外，还展示了我们方法在不同配置特别是在不同的组件和不同的融合方式情况下验证集上的结果。

4.3.1 手工特征

我们从小规模数据集上提取了一些低层特征，包括局部描述子比如：SIFT，HOG，GIST 等。全局颜色直方图（GCH）特征包括 64-bin RGB 直方图，局部颜色直方图特征首先将图片分为 16 个区域，然后对每个区域都使用一个 64-bin RGB 直方图 [45]。我们用 ColorName 去计算图片中 11 种基础颜色中每种颜色的像素，并且用到了 [32]中的算法。我们还用了 SentiBank [3]，一种基于构造本体的概念检测库，将 1200 维特性作为中级表示。Zhao 等 [65]对于情感分析提出了艺术原则的特征。我们用作者提供的简化版本提取了 27 维的特征。

4.3.2 深度模型

You 等 [57]提出的 PCNN 是一个新型渐进的 CNN 架构。他们认为利用大量弱监督数据可以提高模型的普适性。我们用含有噪声的 Flickr 数据集微调了基于

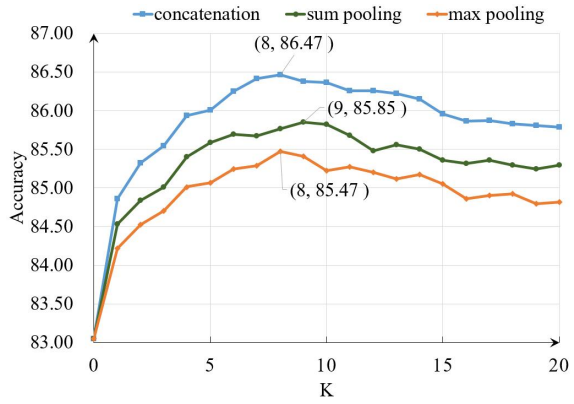

图 7. 不同的 K 值对于 FI 验证集的影响。在剩余的实验中我们设置 $K = 8$。

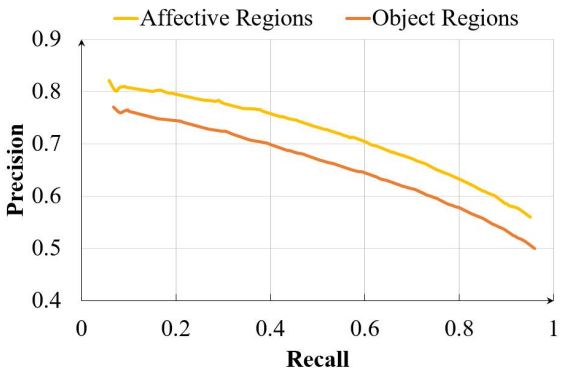

图 8. 发现情感区域的精确率-召回率曲线。和物体区域（用 EdgeBoxes 生成的具有最高物体得分的区域）相比，我们的方法和人工标注更一致。

VGGNet 的 PCNN。DeepSentiBank [7]是基于 CNNs 发现 ANPs 的视觉情绪概念分类。我们用预测训练好的 DeepSentiBank 提取 2089 个 ANPs 作为情感的中层表示。我们还展示了在 ImageNet 上预训练并且在情感数据集上微调过的 CNN 模型深度视觉特征的性能，包括 AlexNet 和 VGGNet 的不同架构。为了和 ImageNet 预训练的 CNN 比较，我们在模型的倒数第二层提取了特征并将其用 PCA 降维，展示了用 LIBSVM [6]在该特征上训练的结果。在实践中，我们发现不同的成本值（LIBSVM 中的参数 C）产生了相似的精度，所以我们只使用默认值，并使用一个 one v.s. all 策略，遵循 [32]中描述的相同的评估方式。

4.4. 在大规模数据集上的结果

我们首先在大规模数据集（FI 和 Flickr）上微调 CNN，并且将我们架构的性能和深度模型相比较。表2报告了 FI 和 Flickr 数据测试集的基准性能。我们可以看到，由于 ImageNet 和情绪数据集分布之间的差异，ImageNet 上的预训练模型不如微调模型，带有更深层次架构的 VGGNet 比 AlexNet 表现得更好。经过微调的 VGG 模型在 FI 数据集上实现了 83.05% 的准确率，超过了 DeepSentiBank（61.54%）和 PCNN（75.34%）。和弱标记的 Flickr 相比，由于可靠的标注，经过微调的 CNN 在 FI 数据集上的表现实现了很大的提升。

当在深度模型选择或合并情感区域时，我们有几个选择：我们可以只用物体得分或者情感得分，或者用该工作提出的 AR 得分。我们粗略地把物体得分看作低级线索，把情感得分看作高级线索。实验结果说明了情感得分比物体得分更有效，主要是因为物体得分只说明了一个区域包含物体的可能性。当两个分数都被合并进深度模型时，我们的方法在置信度最高的区域实现了 84.83% 的准确率，与目前最先进的方法以及只通过一个分数合并目标区域的方法相比表现更好，说明了利用局部细节进行分类的优势。分析不同区域的物体得分和情感得分，我们观察到尽管两个不同的目标区域重合率超过了 50%，但是他们的情感得分总是不同的。对于都包含一个情感区域的两个不同目标区域，他们的情感得分总是相似的，因此可以忽略区域的面积只评估这个目标区域是否包含情感区域。

4.4.1 超参数的影响

我们报告了利用“AR+ 和值池化”的方法在 FI 数据集的验证集上的分类表现，不同的 α 和 β 被用来进行比较。如图 6 所示，令 $\alpha = 0.6$ 实现了在验证集发现情感区域最好的总体精度。只使用物体得分（$\alpha = 0$）得到很一般的结果，说明了在筛选目标区域的过程中考虑情感得分还是很有必要的。另一方面，与只使用单一的全局表示相比，结合局部区域能够促进分类效果。令 $\beta = 0.3$，在大部分情况下实现了平衡。因此，我们在剩余的实验中令 $\alpha = 0.6$，$\beta = 0.3$。

4.4.2 融合操作的影响

当融合情感区域和整个图片的输出时，我们考虑了三种方式去合并置信度最高的情感区域。表2(底部)说明了所有的三个合并方式对于获取全局和局部信息

算法	IAPS-Subset	Abstract	ArtPhoto	Twitter I			Twitter II	EmotionROI
					Twitter I_5	Twitter I_4	Twitter I_3		
GCH	71.76	71.50	67.00	67.91	67.20	65.41	77.68	66.53
LCH	52.91	73.26	64.01	70.18	68.54	65.93	75.98	64.29
ColorName + BoW	57.72	73.28	66.26	64.51	64.79	60.83	70.10	60.13
Gist	65.05	60.97	63.40	65.87	61.47	60.68	77.68	60.38
LBP	56.73	59.85	55.06	55.78	53.94	57.29	65.15	55.26
Gabor	79.21	50.43	58.43	55.37	54.03	53.90	63.72	58.73
SIFT + BoW	86.06	53.54	59.05	63.15	63.71	60.36	70.32	65.30
SIFT + VLAD	83.02	60.53	64.75	70.29	68.91	67.14	77.34	72.15
SIFT + FisherVector	83.28	60.10	62.40	71.09	67.29	65.56	76.34	70.92
DenseSIFT + BoW	56.22	54.38	56.58	64.29	59.94	58.94	60.07	59.85
DenseSIFT + VLAD	58.25	55.74	64.38	67.12	66.49	65.01	77.17	62.13
DenseSIFT + FisherVector	62.55	59.21	64.01	71.76	68.01	65.96	78.01	62.97
HOG + BoW	79.99	60.95	62.40	68.48	61.92	60.99	61.23	61.05
HOG + VLAD	82.52	57.49	68.97	71.99	67.74	66.43	61.92	63.38
HOG + FisherVector	83.76	61.41	68.11	76.07	70.34	68.32	68.12	65.33
PAEF [65]	62.81	70.05	67.85	72.90	69.61	67.92	77.51	75.24
SentiBank [3]	81.79	64.95	67.74	71.32	68.28	66.63	65.93	66.18
DeepSentiBank [7]	85.63	71.19	68.73	76.35	70.15	71.25	70.23	70.11
PCNN (VGGNet) [57]	88.84	70.84	70.96	82.54	76.52	76.36	77.68	73.58
VGGNet	88.51	68.86	67.61	83.44	78.67	75.49	71.79	72.25
Fine-tuned VGGNet	89.37	72.48	70.09	84.35	82.26	76.75	76.99	77.02
obj + 串联连接	88.47	73.38	71.34	84.24	81.81	76.68	75.97	77.83
senti + 串联连接	88.74	74.23	72.86	84.35	82.44	76.57	78.18	77.95
AR + 串联连接	89.39	74.71	73.76	86.10	83.25	77.97	78.89	78.52
AR + 和值池化	90.32	73.72	73.63	86.39	83.41	77.57	78.32	78.43
AR + 最大值池化	89.04	73.92	73.32	86.19	83.11	77.67	78.52	78.32
AR + 串联连接 ($K = 8$)	92.39	76.03	74.80	88.65	85.10	81.06	80.48	81.26

表 3. 不同方法在小规模数据集上的分类结果。GCH 表示全局颜色直方图特征，LCH 对应局部颜色直方图。“obj” 表示我们只将有高物体得分的目标区域作为情感区域，“senti” 表示使用了具有高情感得分的目标区域。

都是有效的，当采用串联连接方式的时候是最有效的方式，因为它保留了全部的信息。

4.4.3 超参数 K 的影响

给定一个输入图片，我们不仅预测整张图片的情感而且还寻找表达情感的区域。尽管数据集没有提供情感区域的标注信息，很多情感区域往往很小。这里我们展示了实验来确定在我们提出的架构中有多少诱发情感的区域。由于缺乏标注，所以直接评估被发现情感区域的质量是比较困难的。因此，我们的目标是调查选择多少情感区域可以提高情感预测的准确率。我们展示了当采用不同数量的情感区域用于情感分析时的分类表现。正如图 7 所示，当情感区域的数量增加时，因为有了更多可用的信息，准确率会增加。然而，进一步增加区域的数量会导致性能方面的略微下降，因为在这个过程中引入了更多的噪声区域。因此，为了更好的平衡，我们在接下来的实验中结合 8 个情感区域用于情感分析，超过了微调的 VGGNet 在 FI（86.35%）

上 3.3%，在 Flickr(71.13%) 上超过了 1%。我们还报告了在大规模数据集上不同情感的真正例率。详细地，在 FI 数据集上，积极的情感和消极的情感分别实现了 92.10% 和 72.65% 的准确率。在 Flickr 上，分别是 73.56% 和 47.92%。对于这两个数据集，积极的类别实现了比消极的类别更高的准确率，这和训练图片的数量是一致的。更多的训练图片会使得相应的情感能有更大的概率获得更高的真正例率。

4.5. 小规模数据集上的结果

我们将在 FI 数据集上学习到的参数迁移到小规模数据集上，然后我们的实验结果展示在了表3上，并且提供了和一些先进方法的比较。我们的方法是基于微调的 VGGNet。“obj” 表示我们只将具有物体得分高的区域作为情感区域，“senti” 指的是具有情感得分高的区域。我们的 “AR” 方法同时考虑物体得分和情感得分进行选择情感区域。

对于颜色特征，与 GCH 和 LCH 相比，除了抽象的数据集，ColorName 通常不足以描述图像颜色的分布。对于纹理特征，HOG 描述子和其他的纹理表示如 SIFT，Gist，LBP 和 Gabor 相比，能够在大部分数据集上实现最好的预测准确率。因为情感常常通过复杂纹理区域表达，比如脸、狗、建筑等。除此之外，我们还比较了表3中不同的编码算法。正如我们看到，当用 Fisher 向量编码这些描述子可在很多数据集上实现更好的表现。

和主要基于颜色和纹理信息的传统表示相比，不出所料深度方法实现了更好的结果。我们提出的方法利用了情感区域，超过了基于手动特征的方法和深度方法，在所有的小数据集上实现了最好的准确率。详细地，和没有利用情感区域而通过中层表示图片的 SentiBank 和 DeepSentiBank 相比，我们的方法强过它们很多。此外，我们的方法在所有的情感数据集上超过了 PCNN，并且三个融合方法都是有用的，其中串联连接是最有效的方法。根据我们在大规模数据集上的实验发现，当我们增加情感区域的数量时许多区域对图像的情感几乎没有影响甚至还降低预测的准确性。因此，我们结合相同数量的区域在小规模的数据集上进行最终的情绪预测并获得最佳性能。这展示了我们方法的另一个优势，我们在深度模型中不需要很多局部区域，确保在计算开销上的增加在可接受范围内。

4.6. 情感区域评估

我们评估该框架在 EmotionROI 数据集上检测到的情感区域，使用了和之前工作 [39, 37]相同的训练/测试划分。由于数据集只提供了标准化的情感刺激图作为正确标签，它基于 15 个边界框，我们首先用阈值 $\gamma \in [0..255]/255$ 将情感刺激图二值化，并将标注的区域与发现的置信度最高的情感区域进行比较。我们还使用了精确率和召回率，代表了在预测区域或标注区域内识别出的所有像素中所检测到的情感相关像素的百分比。依据 [39]，为了评估便捷，所有预测到的情感区域和标注区域都被正则化到 0 到 1。图 8 展示了物体得分和 AR 得分的精确率和召回率曲线。我们方法的平均精确率和召回率分别是 0.69 和 0.59，在这两项指标上，物体得分的度量分别实现了 0.63 和 0.53，说明了我们所选的情感区域与人类的标注更一致。

5. 情感区域的可视化

对于图片分类的方式，一个自然的问题是所提出的模型能否识别图像中的目标部分。在本节中，我们尝试通过可视化对于情感分类重要的区域来回答这个问题。根据之前的工作 [61]，我们用灰色滑动窗口遮住输入图片的不同位置，然后通过描绘对应位置真实类别的估算概率生成热力图。和其他可视化方式相比，比如用 t-SNE 对特征进行映射或者可视化网络的滤波器，这些方法倾向于直接显示 CNN 关注的区域。如图 9 所示，第一列是输入图片，第二列是当遮挡图片中对应位置用微调过的 VGGNet 预测出的正确类别的概率，如果遮挡的位置对于情感预测比较重要，那么在热力图中对应的概率将会明显下降（蓝色像素）。正如可以看到的三个例子，微调过的深度模型有能力发现图片中诱发情感的区域。比如，遮住可以诱发情感的显著物体（比如人，火）将会降低预测的概率。然而，由于情感鸿沟，卷积神经网络没有足够的区分能力去获取在图片中最重要的情感信息。

分别根据图 9 (c)(d)（e）中不同的得分（物体得分、情感得分、AR）我们重排了候选区域，并可视化了排名第一的区域。列 (c) 和列 (d) 指的只根据物体得分或情感得分选择的区域。根据物体得分，选择了在低层包含丰富信息的区域，而根据情感得分，则往往对情感层面上对区域情感进行评估。当同时考虑来自这两方面

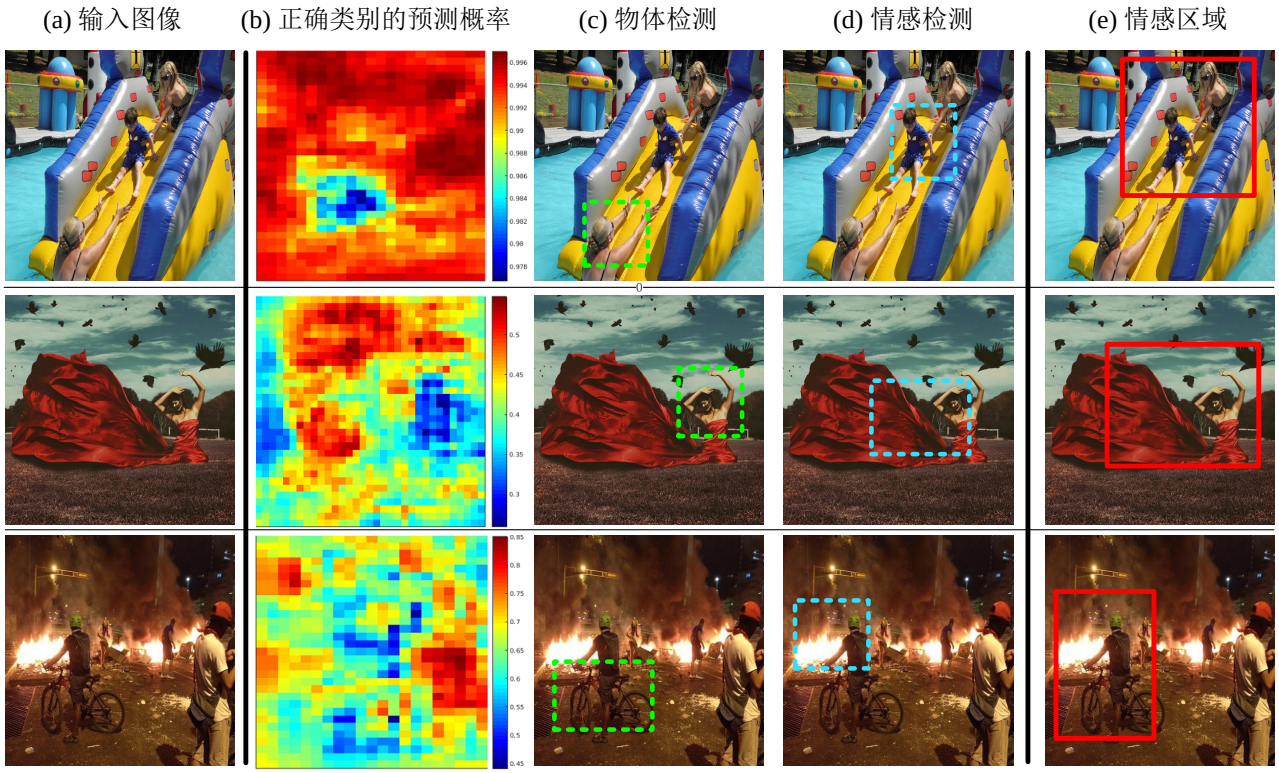

图 9. FI 数据集中图片的可视化。给定输入图片 (a)，我们有组织地用灰色方块遮挡图片的不同位置，观察分类器的输出 (b) 怎样改变。列 (b) 表示用 CNN 对正确类别的概率估计图，表明了对于 CNN 来说情感图片中位置的相对重要性。我们还展示了通过不同得分 (Obj_score,Senti_score,AR) 进行排序后排名第一的区域，作为物体区域 (c)，情感检测区域 (d)，情感综合区域 (e)。

的信息，我们提出的方法能够发现更准确的情感区域，由列 (e) 看出。检测到的情感区域不仅可以对图像中的显著物体进行补充（第一个例子），还可以将感兴趣的区域扩展到额外的上下文背景（最后两个例子）。因此，将全局信息和局部信息结合对于视觉情感分析是有识别力的。

6. 结论

在本文中，我们解决了自动识别图片中情感的问题。受到全局表现和局部区域都能产生显著情感反应的启发，我们提出了一个框架去发掘情感区域并可以用卷积神经网络将全局特征和局部特征结合。然后依据物体得分和情感得分估计一个区域内情感内容的水平。通过物体得分往往会发现包含丰富纹理信息的区域，而情感得分会在情感层面评估区域的情感。我们还考虑了三种可选择的融合操作，并在 VGGNet 的基础上实现了这个模型。实验结果显示我们的方法在主流情感数据集上的表现超过了目前最先进的方法。

致谢 本文工作由 NSFC (NO. 61620106008,61572264, 61633021, 61525306, 61301238, 61201424), theOpen Project Program of the National Laboratory of PatternRecognition (NLPR), Huawei Innovation Research Program,CAST YESS Program, and IBM Global SUR award 支持。

参考文献

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell., 34(11):2189–2202, 2012. 5

[2] S. Benini, L. Canini, and R. Leonardi. A connotative space for supporting movie affective recommendation.

IEEE Trans. Multimedia, 13(6):1356–1370, 2011. 2

[3] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In ACM Int. Conf. Multimedia, 2013. 1, 2, 3, 5, 7, 9, 11

[4] V. Campos, B. Jou, and X. Giró-I-Nieto. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. Image Vision Comput., 65:15–22, 2017. 1, 3

[5] V. Campos, A. Salvador, X. Giró i Nieto, and B. Jou. Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction. In International Workshop on Affect & Sentiment in Multimedia, 2015. 1, 3, 8

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Trans. Intel. Syst. Tec, 2(3):1–27, 2011. 10

[7] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. ArXiv e-prints, 2014. 3, 9, 10, 11

[8] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang. Object-based visual sentiment concept analysis and application. In ACM Int. Conf. Multimedia, 2014. 2, 3, 4, 5

[9] Y.-Y. Chen, T. Chen, T. Liu, H.-Y. M. Liao, and S.-F. Chang. Assistive image comment robot—a novel mid-level concept-based representation. IEEE Trans. Affect. Comput., 6(3):298–311, 2015. 1

[10] M. Cheng, Z. Zhang, W. Lin, and P. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In IEEE Conf. Comput. Vis. Pattern Recog., 2014. 5

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recog., 2009. 3

[12] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. Annu. Rev. Neurosci., 18(1):193–222, 1995. 2

[13] E. Elhamifar and R. Vidal. Sparse subspace clustering. In IEEE Conf. Comput. Vis. Pattern Recog., 2009. 5

[14] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In Int. Conf. Comput. Vis., 2015. 4

[15] R. Girshick. Fast R-CNN. In Int. Conf. Comput. Vis., 2015. 4

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conf. Comput. Vis. Pattern Recog., 2014. 3, 4, 5

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 38(1):142–158, 2016. 4

[18] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with R*CNN. In Int. Conf. Comput. Vis., 2015. 4

[19] A. Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. IEEE Signal Proc. Mag., 23(2):90–100, 2006. 2

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conf. Comput. Vis. Pattern Recog., 2016. 1

[21] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? IEEE Trans. Pattern Anal. Mach. Intell., 38(4):814–830, 2016. 5

[22] J. H. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In Brit. Mach. Vis. Conf., 2014. 5

[23] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. IEEE Signal Proc. Mag., 28(5):94–115, 2011. 1

[24] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated GIFs. In ACM Int. Conf. Multimedia, 2014. 2

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Adv. Neural Inform. Process. Syst., 2012. 3, 6, 9

[26] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report, 2008. 7

[27] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural. Comput., 1(4):541–551, 1989. 3

[28] B. Li, W. Xiong, W. Hu, and X. Ding. Context-aware affective images classification based on bilayer sparse representation. In ACM Int. Conf. Multimedia, 2012. 2, 4

[29] Z. Li, Y. Fan, W. Liu, and F. Wang. Image sentiment prediction based on textual descriptions with adjective noun pairs. Multimed. Tools Appl., pages 1–18, 2017. 3

[30] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. RAPID: Rating pictorial aesthetics using deep learning. In ACM Int. Conf. Multimedia, 2014. 1, 5

[31] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang. On shape and the computability of emotions. In ACM Int. Conf. Multimedia, 2012. 1, 3

[32] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In ACM Int. Conf. Multimedia, 2010. 1, 2, 3, 7, 9, 10

[33] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. Behavior Research Methods, 37(4):626–630, 2005. 3, 7, 8

[34] M. A. Nicolaou, H. Gunes, and M. Pantic. A multilayer hybrid framework for dimensional emotion classification. In ACM Int. Conf. Multimedia, 2011. 2

[35] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. Found. Trends Inf. Ret., 2(1–2):1–135, 2008. 1

[36] L. Pang, S. Zhu, and C. Ngo. Deep multimodal learning for affective analysis and retrieval. IEEE Trans. Multimedia, 17(11):2008–2020, 2015. 1

[37] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In IEEE Conf. Comput. Vis. Pattern Recog., 2015. 12

[38] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In IEEE Int. Conf. Image Process., 2016. 2, 4, 7

[39] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In IEEE Int. Conf. Image Process., 2016. 3, 9, 12

[40] X. Peng and C. Schmid. Multi-region two-stream R-CNN for action detection. In Eur. Conf. Comput. Vis., 2016. 4

[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In IEEE Conf. Comput. Vis. Pattern Recog., 2016. 4

[42] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Adv. Neural Inform. Process. Syst., 2015. 2, 4

[43] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe. Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings. In ACM Int. Conf. Multimedia, 2015. 1

[44] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 22(8):888–905, 2000. 5

[45] S. Siersdorfer, E. Minack, F. Deng, and J. Hare. Analyzing and predicting sentiment of images on the social web. In ACM Int. Conf. Multimedia, 2010. 9

[46] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Int. Conf. Learn. Represent., 2015. 5, 6, 8, 9

[47] M. Solli and R. Lenz. Color based bags-of-emotions. In Int. Conf. Comput. Anal. Images Patterns, 2009. 2

[48] M. Sun, J. Yang, K. Wang, and H. Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In Int. Conf. Multimedia and Expo, 2016. 2

[49] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders. Segmentation as selective search for object recognition. In Int. Conf. Comput. Vis., 2011. 5

[50] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. HCP: A flexible CNN framework for multi-label image classification. 38(9):1901–1907, 2016. 5

[51] W. Weining, Y. Yinglin, and J. Shengming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In Int. Conf. Syst. Man. Cy., 2006. 7

[52] H. Wu, H. Zhang, J. Zhang, and F. Xu. Typical target detection in satellite images based on convolutional neural networks. In Int. Conf. Syst. Man. Cy., 2015. 5

[53] J. Yang, D. She, and M. Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In Int. J. Conf. Artif. Intell., 2017. 3

[54] J. Yang, M. Sun, and X. Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In AAAI Conf. Artif. Intell., 2017. 3

[55] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek. Emotional valence categorization using holistic image features. In IEEE Int. Conf. Image Process., 2008. 1, 7

[56] Q. You, H. Jin, and J. Luo. Visual sentiment analysis by attending on local image regions. In AAAI Conf. Artif. Intell., 2017. 2, 3

[57] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In AAAI Conf. Artif. Intell., 2015. 1, 2, 3, 7, 8, 9, 11

[58] Q. You, J. Luo, H. Jin, and J. Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In AAAI Conf. Artif. Intell., 2016. 1, 3, 7

[59] Q. You, J. Luo, H. Jin, and J. Yang. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In ACM Int. Conf. Web Search and Data Mining, 2016. 1

[60] J. Yuan, S. Mcdonough, Q. You, and J. Luo. Sentribute: image sentiment analysis from a mid-level perspective. In ACM International Workshop on Issues of Sentiment Discovery and Opinion Mining, 2013. 3

[61] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Eur. Conf. Comput. Vis. 2014. 12

[62] L. Zhang, L. Lin, X. Liang, and K. He. Is faster R-CNN doing well for pedestrian detection? In Eur. Conf. Comput. Vis., 2016. 4

[63] S. Zhao, G. Ding, Y. Gao, and J. Han. Approximating discrete probability distribution of image emotions by multi-modal features fusion. In Int. J. Conf. Artif. Intell., 2017. 3

[64] S. Zhao, Y. Gao, G. Ding, and T. S. Chua. Real-time multimedia social event detection in microblog. IEEE Trans. Cyber., PP(99):1–14, 2017. 1

[65] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In ACM Int. Conf. Multimedia, 2014. 1, 3, 7, 9, 11

[66] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua. Predicting personalized image emotion perceptions in social networks. IEEE Trans. Affect. Comput., 2018. 2

[67] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. IEEE Trans. Multimedia, 19(3):632–645, 2017. 3

[68] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua. Predicting personalized emotion perceptions of social images. In ACM Int. Conf. Multimedia, 2016. 2

[69] S. Zhao, H. Yao, and X. Jiang. Predicting continuous probability distribution of image emotions in valence-arousal space. In ACM Int. Conf. Multimedia, 2015. 3

[70] S. Zhao, H. Yao, Y. Yang, and Y. Zhang. Affective image retrieval via multi-graph learning. In ACM Int. Conf. Multimedia, 2014. 3

[71] L. Zheng, Y. Yang, and Q. Tian. SIFT meets CNN: a decade survey of instance retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 2017. 1

[72] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In Adv. Neural Inform. Process. Syst., 2014. 1

[73] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In Eur. Conf. Comput. Vis., 2014. 4, 5