深度学习结构优化的图像情感分类

盛家川[1,2], 陈雅琦[1], 王　君[3], 韩亚洪[4*]

(1. 天津财经大学 理工学院，天津 300222；
2. 金融科技与风险管理实验室，天津 300222；
3. 天津财经大学 管理科学与工程学院，天津 300222；
4. 天津大学 智能与计算学部，天津 300072)

摘　要：自然图像情感分类在分析用户需求、监控网络舆情等方面具有重要意义。然而基于深度学习的分类算法存在训练过程难以控制、分类结果缺乏解释的问题。为此提出一种人类知识驱动的深度学习结构优化算法。首先通过特征可视化显示卷积神经网络提取的情感特征;其次结合人类对图像情感可视化结果的感知来优化网络结构,利用人类知识驱动网络,重点学习情感信息更明显的特征;最后对所构建网络的参数进行微调,使其更适用于自然图像情感分类任务。在 Twitter 情感图像数据集上与其他分类方法的对比实验表明,所提出的算法获得了 88.1% 的分类准确率,优于其他方法。消融实验证明网络优化结果比未优化提高了 8.1%。类激活图、空间位置和神经元组特征可视化直观解释了模型运作的过程与原因,进一步证实算法识别自然图像情感的能力。

关键词：图像情感；　图像分类；　卷积神经网络；　可视化；　类激活图
中图分类号：TP391.4　　**文献标志码：**A　　**DOI：**10.3788/IRLA20200269

Image sentiment classification via deep learning structure optimization

Sheng Jiachuan[1,2], Chen Yaqi[1], Wang Jun[3], Han Yahong[4*]

(1. School of Science and Technology, Tianjin University of Finance &Economics, Tianjin 300222, China;
2. Laboratory of Fintech and Risk Management, Tianjin 300222, China;
3. School of Management Science and Engineering, Tianjin University of Finance & Economics, Tianjin 300222, China;
4. College of Intelligence and Computing, Tianjin University, Tianjin 300072, China)

Abstract: Automatically analyzing the sentiment of natural images plays a vital role in analyzing user needs and network public opinion monitoring. However, the training processes of deep learning-based classification algorithms are too difficult to be controlled, and their classification results are always lack of interpretation. A deep learning structure optimization algorithm with human cognition was proposed to classify image sentiment. Firstly, the emotional features extracted were visualized by the convolutional neural networks. Then, the network structure was optimized by combining with human's subjective perception of image emotion, and the network structure was driven by human knowledge to focus on the apparent features of emotional information. Finally, the parameters of the rebuilt network were fine-tuned to make it more suitable for images sentiment classification

收稿日期:2020−06−11；　**修订日期:**2020−07−15
基金项目:国家自然科学基金 (61502331, 61876130); 教育部人文社科项目 (18YJA630057, 19YJA630046); 天津市自然科学基金 (18JCYBJC85100); 天津市企业科技特派员项目 (19JCTPJC56300)
作者简介:盛家川 (1982−), 女, 教授, 硕士生导师, 博士, 主要从事多媒体处理、模式识别方面的研究工作。Email: jiachuansheng@tjufe.edu.cn
通讯作者:韩亚洪 (1977−), 男, 教授, 博士生导师, 博士, 主要从事多媒体分析、计算机视觉和机器学习的研究工作。Email: yahong@tju.edu.cn

task. Contrastive experiments on the Twitter dataset systematically demonstrate that the proposed algorithm achieves 88.1% classification accuracy, which has superior performance than other methods. Ablation experiments confirm that our network optimization improves the classification effect by 8.1%. Besides, the process and reason were intuitively explained for the model operation through class activation maps, spatial location visualization and neuron group visualization. The visualization experimental results further demonstrate the ability of proposed algorithm to recognize the sentiment of natural images.

Key words: image sentiment; image classification; convolutional neural networks; visualization; class activation map

0 引 言

随着互联网的发展以及移动终端设备的普及，越来越多的人热衷于将自己的文字与图片分享到微博、Twitter 等社交平台上。“一幅图胜过千言万语”，与文本信息相比，图像蕴含了更加丰富、抽象的信息，且图像跨越了语言的障碍，具有全球通用性。网络上图片内容多种多样，而图像所表达的情感语义信息在很大程度上传达了用户的价值观和心理活动。通过研究海量图像的情感，不仅能够了解用户情感需求，在图像检索、推荐系统等方面帮助优化用户体验，而且有助于研究社会热点问题，获悉大众观点态度倾向，为网络舆情分析与监控提供数据支持。

图像情感分类研究起步相对较晚。有些研究通过提取图像底层特征，结合心理学知识或视觉认知分析图像情感语义。Yao 等[1] 提取人脸表情动作单元特征并寻找它们之间的关系实现视频的情感识别。Kang 等[2] 构建颜色组合和情感词数据集，根据为图像生成的颜色谱搜索数据集中最佳匹配的情感词。随着卷积神经网络 (Convolutional Neural Network, CNN) 被广泛地应用于各种图像分类、识别等任务，且展现出优异的性能，深度学习逐渐被引入图像情感分类的研究工作中。Song 等[3] 将显著图引导的视觉注意力机制与卷积神经网络架构结合，取得较好的情感分类性能。He 等[4] 使用二分类 CNN 网络的结果来辅助识别图像的多种情感，但需要为数据集重新赋二分类所需的情感标签。

现有基于深度学习的图像分类研究多使用数据驱动的方法，而知识驱动却很少受到重视。当前研究存在网络决策缺乏解释、模型内部学习过程难以控制的缺点[5−6]。近年来一些文献利用可视化技术解释 CNN 决策，但没有利用可视化结果进行基于知识驱动的图像分类。Fu 等[7] 可视化并分析预训练网络不同层次在输入图像中的关注焦点，对比了 AlexNet 和 VGG 模型识别对象位置的能力。Olah 等[8−9] 利用特征可视化，显示在 ImageNet 上训练的 GoogLeNet 所检测到的特征，解释网络如何理解不同的物体对象并分辨它们。类激活映射 (Class Activation Mapping, CAM) 技术[10−11] 能够根据类别标签显示图像不同区域对最终分类的作用大小，且与其他可视化算法[12−13] 相比计算量小、定位效果明显并易于理解。以上可视化技术仅在模型训练完成后起解释作用，人类从可视化结果获得的知识并未对 CNN 的架构做出改进。

图像情感本质上是图像引发人类产生的心理反应，因此如果将人对情感的认知注入分类网络，则有助于增强系统可控性，提高识别能力。针对以上问题，文中利用深度网络特征可视化技术帮助理解网络学习情感信息的过程，进而构建更适应自然图像情感分类任务的网络模型。通过可视化网络内部特征将人类知识与深度学习技术结合起来，对推动人工智能发展，实现人机结合具有重要意义。

文中贡献主要包括量个方面：(1) 提出基于 CNN 特征可视化的互动式网络架构优化方法，结合数据驱动与知识驱动，通过人类对图像情感语义表达的主观感知，优化了针对自然图像情感分类任务的 CNN 网络结构；(2) 将网络学习到的情感特征可视化为人能够理解的图像，利用通道可视化和神经元组可视化为模型的情感决策生成“视觉解释”，使网络训练符合人类认知图像情感的方式，为系统优化决策提供有力依据。

1 相关工作

卷积神经网络在特征提取及分类方面具有明显

的优势。深层网络结构能够获取更抽象的高层语义特征，从而处理较为复杂的分类任务[14−20]。Szegedy 等在 ILSVRC14 大赛上提出 22 层深的卷积神经网络模型 GoogLeNet 并获得了最优成绩[21]。GoogLeNet 创新性地由 Inception 模块构建而成，在增加了网络的宽度和深度的同时减少了训练参数。Inception 模块输出与分类器之间使用了全局平均池化，保留了二维特征信息与类别之间的关联。由于这些特点，GoogLeNet 相较于 VGG[22] 等网络不仅卷积层尺寸大小相对固定，方便不同层之间可视化图的对比，而且其可视化更具语义意义[9]，有助于人类肉眼观察理解，为主观感知注入模型奠定基础。

GoogLeNet 通过多层的非线性变换，训练得到有助于分辨数据集类别的高级特征。网络中间使用了最大池化层，在识别物体时最大池化能够获取对象最突出的特征，提高模型准确性，却也丢失了大量信息。而在辨别情感时不只需要某些突出特征，基本形状、物体对象、整体氛围等因素同样对图像情感渲染非常重要。因此 GoogLeNet 这种抛弃部分网络中间特征的训练方式不利于图像的情感分类。另外，GoogLeNet 虽然使用辅助分类器辅助训练，但在 ImageNet 分类任务中作用较小，且辅助分类器并未参与结果预测。

2 人类知识驱动的图像情感分类网络优化

在并不清楚网络具体提取了什么特征的情况下，用户仅能盲目地获取结果。只有真正了解 CNN 训练提取特征的运作过程，才能针对性地调整网络学习策略，有效干预网络学习过程，使计算机的判断更符合人类的预期。文中引入可视化协助人类观察理解网络各层究竟学习到了怎样的特征，在这一过程中通过人的知识和心理感受，来判断卷积层训练的特征对于情感分类任务的重要程度，进而优化网络结构，重点学习辨别的情感所需特征，控制网络训练方向，提高特征利用率。算法过程如图 1 所示。文中使用自然图像情感数据集微调 GoogLeNet(上方标注卷积层输出尺寸)，并可视化网络学习到的通道和神经元组特征。人类通过观察特征，优化深度学习网络结构，结合主观情感认知选择对情感分类更有帮助的卷积层，为其添加辅助分类器 (红色线框标明) 引导网络训练方向。最终分类结果由所有分类器的加权输出决定。

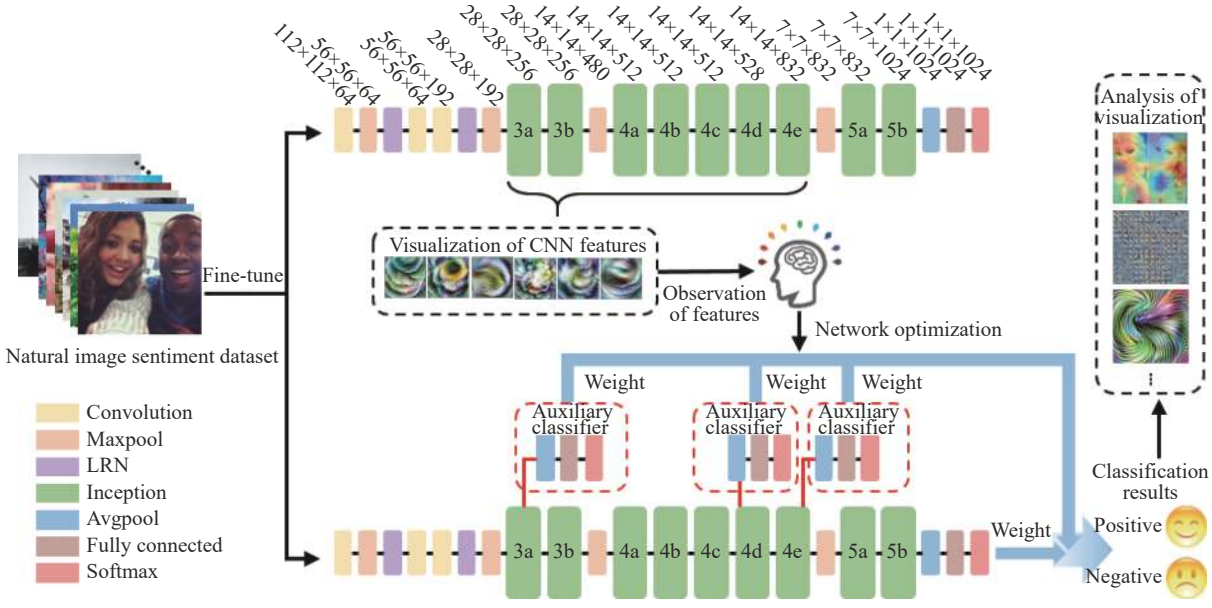


图 1 文中算法过程

Fig.1 Overview of proposed algorithm

2.1 网络可视化分析

黑格尔曾说：“情感是心灵中的不确定的模糊隐约的部分。”情感是一种高度抽象的信息，有时甚至用语言都很难形容。对于图像情感这样主观而又抽象的信息，更需要剖析深度网络提取特征并做出判断的内部过程及原因。特征可视化可以将 CNN“黑盒”

打开满足研究者的好奇心，赋予网络可解释性，帮助优化网络结构，并进一步证明模型的情感感知能力，有助于模型的推广应用。

如图 2 所示，文中引入激活最大化 (Activation Maximization) 技术[23] 从卷积层通道、空间位置、神经元组三个角度可视化情感特征，其中红色区域表示被可视化的特征。一方面，通过可视化实现人类情感知识与网络内部信息的互动，借助人类对图像情感的主观感受构建更适用于情感分类的网络。另一方面，深入理解训练得到的自然图像情感分类网络的工作原理，直观解释网络决策原因，提高算法可解释性。

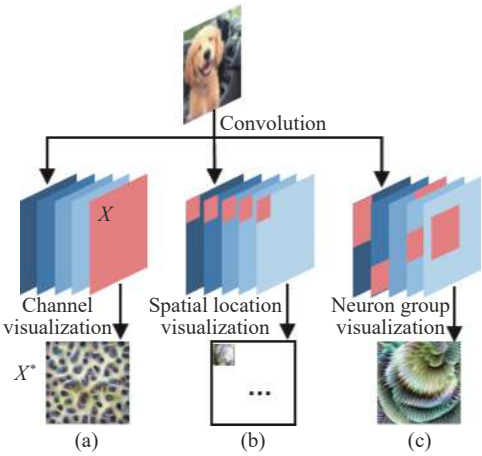


图 2 卷积层特征可视化

Fig.2 Features visualization of the convolutional layers

2.1.1 通道可视化分析

激活最大化技术目标是找到使得激活响应最大的输入。若一个输入图像能够使某个通道特征激活高，则图像包含了该通道所表示的特征。但符合条件的图像数量较大，且图中包含了多种特征，仅仅将它们列出来很难对单独的特征进行观察。激活最大化技术通过使一个通道的激活最大来生成可视化图像。该图像可以最大程度地总结相应卷积核从数据集学习到的某种特征，同时屏蔽其他特征的干扰。

在模型训练结束的情况下，网络参数θ已经确定。设$h_{ij}(\theta,X)$为网络第 j 层中单元 i 的激活，h_{ij}是参数θ(包括权重和偏差) 和输入样本 X 的函数，则通道特征可视化能够转化为寻找输入 X*：

$$X^*=\arg\max_{X\ \mathrm{s.t.}\|X\|=\rho} h_{ij}(\theta,X) \tag{1}$$

式中：X*表示当$h_{ij}(\theta,X)$最大时 X 的取值，即通道特征可视化。这是一个非凸优化问题，可以通过梯度上升法求得局部最大值。具体来说，首先用随机值初始化输入图像 X，接着计算$h_{ij}(\theta,X)$的梯度$\frac{\partial h_{ij}}{\partial X}$，设置适当的学习速率并在梯度方向上调整输入图像 X。设当前迭代的可视化图像为 X'，下一次迭代结果为图像 X''，则一次梯度上升迭代更新的计算如下：

$$X''=X'+\alpha\frac{\partial h_{ij}}{\partial X} \tag{2}$$

式中：α表示学习速率。图像 X 的像素在每次迭代后作出改变，最终逐渐拟合通道特征得到输出的通道可视化结果 X*。文中用正态分布随机初始化 224×224×3 的矩阵 (尺寸与卷积网络输入图像一致)，将其作为输入 X。实验中取学习速率$\alpha=0.05$训练，以获得较为清晰的可视化图像。图 2(a) 显示了卷积层的通道可视化，卷积层的每一个通道可以被可视化为一种特征。

2.1.2 空间位置可视化分析

下面从空间位置角度组合神经元特征，尽可能将网络提取到的自然图像情感特征可视化为人类可以理解的图像。由于神经元可视化通常只能代表很少几个样本中的特征，可视化意义较小。而通道可视化表现更加全面，也就是说神经元可视化是通道可视化的局部片段。因此，文中使用通道可视化替代神经元可视化作为基础计算组合特征。

设通道尺寸为 $m\times n$，该层卷积核尺寸为 $l\times l$，将通道分为$\frac{m}{l}\times\frac{n}{l}$个位置格。设$p_{ij}^k$为第 k 通道第 i 行、第 j 列位置神经元的特征可视化，则P^k为通道 k 的特征可视化。图像不同位置激活大小不同，即图像不同空间位置神经元特征组合不同。设w_{ij}^k为p_{ij}^k对应权重。空间位置特征图 Q 的第 i 行、第 j 列位置格的可视化图像q_{ij}可表示为：

$$q_{ij}=\sum_{k=1}^{t}P^k w_{ij}^k \tag{3}$$

式中：t 表示该卷积层通道数量。空间位置特征即一个卷积层相应位置所有神经元特征的加权和。可视化固定位置激活向量提取的特征，能够帮助分析和理解图像不同位置纹理、对象的特征信息。如图 2(b) 所示，利用公式 (3) 分别计算每一个位置的可视化图像，即可得到一幅图像的空间位置可视化特征图。

2.1.3 神经元组可视化分析

空间位置可视化和通道可视化简单地在卷积层

上纵向或者横向组合神经元，因此可视化结果只关注了特征的一个方面——图像位置或者单独一种特征模式。文中应用非负矩阵分解获取神经元组，将卷积层中高度相关的一系列神经元视为一个整体，找到更有意义的神经元组合方式，使可视化图像容易被观察者理解。非负矩阵分解与人类认识事物的过程相似，反映了“局部构成整体”的概念，分解得到的向量具有语义意义。

用 B 代表卷积层激活，设该卷积层有 t 个尺寸为 $m\times n$ 的通道，那么 B 是 $m\times n\times t$ 维矩阵。由于修正线性单元 (Rectified Linear Unit, ReLU) 激活后所有元素都是非负的，满足非负矩阵分解要求。B 近似分解为基矩阵 U 和系数矩阵 V：

$$B_{m\times n\times t}\approx U_{m\times n\times r}V_{r\times t} \tag{4}$$

式中：r 表示分解的神经元组的数量，远小于 m、n 和 t。鉴于只需观察图像积极或消极中的一种情感，文中取 r=1，则得到 t 维系数向量 V。神经元组可视化 S 即以 V 作为权重系数的通道特征加权和：

$$S=\sum_{k=1}^{t}P^{k}v_{k} \tag{5}$$

神经元组可视化示意如图 2(c) 所示，利用非负矩阵分解得到某个卷积层高度相关的神经元，将这些神经元特征可视化，并以公式 (4) 分解获取的向量 V 作为权重，用公式 (5) 进行加权组合，得到的可视化结果与图像输出类别存在着更为紧密的联系，语义信息更强。

2.2 基于人类情感认知的网络结构优化

特征可视化能够帮助研究者直观地了解深层网络工作原理。为了观察网络获得的情感特征，文中将自然图像情感数据集作为输入，微调在 ImageNet 上预训练的 GoogLeNet 模型。利用 ImageNet 预训练的模型对自然图像情感进行特征提取是可行的。由于 ImageNet 和文中图像情感数据集同为自然图像，参与训练的图像有一千多万张，ImageNet 模型包含十分丰富的纹理、颜色、物体对象等特征，直接或间接包含自然图像情感相关的特征，因此有助于对图像情感的分类。微调能够弥补数据集较小的缺陷，减少训练时间，使训练所得的特征蕴含情感信息。文中将 GoogLeNet 中的 Inception 模块组合视为一个整体层，通过可视化卷积层的通道特征和神经元组特征观察各层特征。

CNN 的一个卷积层由许多通道组成，每个通道代表一种特征。图像特征正是由这些特征加权组合而成，因此可视化模块的通道特征可以辅助理解网络各层特征的模式。如图 3 所示，网络不同深度的通道学习特征角度不同，浅层倾向于提取物体局部纹理、物体对象形状等特征；而深层倾向于提取图像整体氛围信息，可视化图像十分复杂抽象。由于人类会结合自身知识对图片情感进行推理判断，其感受到的情感往往会受到简单形状、物体目标的象征意义的影响。例如矩形给人稳固、安全有序之感，三角形表现紧张与侵略性；植物嫩芽体现勃勃生机，一般表达积极情感；枪支通常代表战争、暴力，表现负面情感。CNN 的特征信息从浅层向深层逐层传递，然而，其中使用的最大池化层虽然能够减少参数量，获取感受野区域的突出特征，同时却导致一些中间层信息的丢失。因此增强中间层简单特征对模型训练的影响有助于图像情感的识别。

微调后的模型中间 7 层 Inception 模块的神经元组特征可视化如图 4 所示，其中左侧为正类图像神经元组可视化，右侧为负类图像神经元组可视化。由于非负矩阵分解得到了最相关的一组神经元，所以不同情感类别图像的神经元组可视化差异越大，表示该卷

(a) 浅层
(a) Shallow layers

(b) 深层
(b) Deep layers

图 3 微调模型通道可视化示例

Fig.3 Channel visualization example of the fine-tuned model

积层的特征越针对情感信息。观察并对比大量可视化图像可知，Layer (4a)、Layer (4b)、Layer (4d) 层正类图像相较于负类图像，其神经元组可视化的线条、颜色分布更有规律，两类图像的神经元组可视化图像区别更为明显，直观上该三层学习到的情感信息更加丰富，与其他卷积层相比更能辨别积极或消极情感的特点，有助于情感的辨别。文中选取特征类别差异明显的 Layer (4a)、Layer (4b)、Layer (4d) 层截取输出特征，添加辅助分类器并按权重参与到模型训练与分类中，使得人类感受到的情感因素参与计算机识别，引导网络重点训练情感相关的特征，增加辨别情感能力较强层的特征利用率，同时起到减轻过拟合的作用。

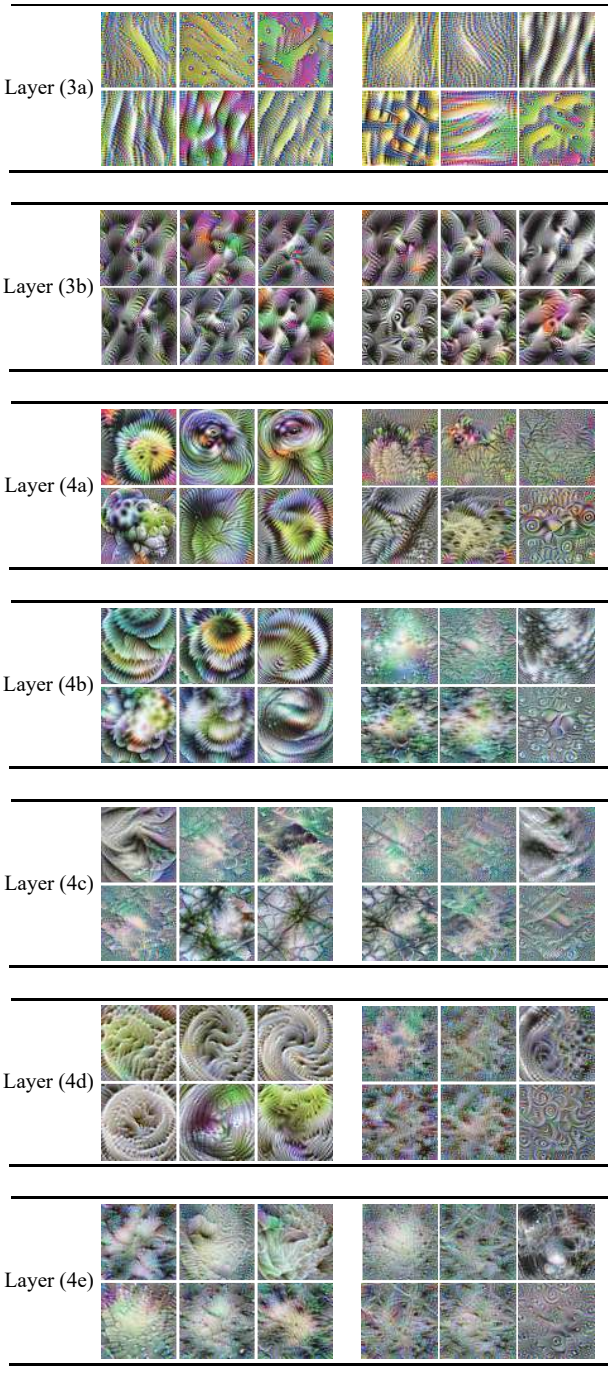


图 4 微调模型神经元组可视化

Fig.4 Neuron group visualization example of the fine-tuned model

辅助分类器使用全局平均池化代替全连接层映射得到中间层特征向量，并用 Dropout 缓解过拟合。全局平均池化层强化了通道特征与特征向量的对应关系[24]，即全局平均池化层得到的特征向量中的元素代表一个通道所提取的特征，因而辅助分类器结果可以充分体现人类认知的情感因素。文中尽量简化辅助分类器网络构造，从而在防止过拟合的同时，最大程度保留中间层信息。其中 3 个辅助分类器详细信息见表 1。

表 1 辅助分类器详细信息

Tab.1 Details of auxiliary classifier

	Global average pooling		Output size of dropout	Output size of full connection	Output size of Softmax
	Kernel size	Output size			
Auxiliary classifier of Layer (4a)	14×14	1×1×512	1×1×512	1×1×2	1×1×2
Auxiliary classifier of Layer (4b)	14×14	1×1×512	1×1×512	1×1×2	1×1×2
Auxiliary classifier of Layer (4d)	14×14	1×1×528	1×1×528	1×1×2	1×1×2

3 实验结果与分析

3.1 实验与数据分析

使用 Twitter 数据集[25] 验证文中算法。该数据集中的每个图像都由 5 人标注情感标签 (正类或负类)，详细信息如表 2 所示。其中 five agree 指 5 人为图像提供了相同的标签，four agree 指至少 4 人标注的标签是相同的。即 five agree 集标注质量最高，three agree 集噪声最大。

预处理阶段，为了提取更详细的情感特征，通过中心裁剪调整图像大小，并随机翻转图像，有助于避免过拟合问题，提高模型的泛化能力。保留 GoogLeNet 在 ImageNet 数据集上预训练的权重，随机初始化输

表 2 Twitter 数据集详细信息

Tab.2 Details of the Twitter dataset

	Five agree	Four agree	Three agree
Positive	581	689	769
Negative	301	427	500
Total	882	1 116	1 269

出层和辅助分类器的权值，用 Twitter 自然图像情感数据集微调文中构建的结合人类认知的自然图像情感分类网络，既可以借助 ImageNet 自然图像特征使得网络收敛更快，又能减轻数据集较小情况下的过拟合问题。其中 Layer (4a)、Layer (4b)、Layer (4d) 层辅助分类器分别以 0.1 的权重参与模型训练与分类。

采用 5 折交叉验证来评估模型性能。分别计算 Five agree、Four agree、Three agree 三个数据集测试结果的精度 (*P*)、召回率 (*R*)、*F*1 分数、准确率 (*A*)4 个评估指标，与 PCNN[25] 和 SentiNet-A[3] 算法实验结果对比如表 3 所示。评估指标对比结果表明，文中提出的算法获得了更高的准确率。

接下来通过对不同结构的网络在 Five agree 数据集上的五折交叉验证来证明文中模型质量。分别在网络各层添加辅助分类器，使其以 0.1 的权重参与模型训练与分类，探究各层辅助分类器对分类的影响，结果如表 4 所示。实验 (1) GoogLeNet 不添加任何辅助分类器获得了 80% 的准确率。实验 (2)~(5) 显示，在 Layer (3a)、Layer (3b)、Layer (4c)、Layer (4e) 层添加辅助分类器对分类结果影响很小，其他 3 层添加辅助分类器实验结果见表 5 中的实验 (5)~(7)。以上实验证明，通过可视化的确能够辨别卷积层中的情感信息。

表 3 对比实验结果

Tab.3 Results of comparative experiment

		PCNN[25]	SentiNet-A[3]	Proposed algorithm
Five agree	*P*	0.770	0.895	**0.900**
	R	0.878	0.878	**0.922**
	*F*1	0.821	0.886	**0.911**
	A	0.747	0.851	**0.881**
Four agree	*P*	0.733	0.851	**0.860**
	R	0.845	0.835	**0.877**
	*F*1	0.785	0.842	**0.868**
	A	0.714	0.807	**0.836**
Three agree	*P*	0.714	0.823	**0.836**
	R	0.806	0.809	**0.848**
	*F*1	0.757	0.814	**0.842**
	A	0.687	0.777	**0.807**

注：加粗字体为每行最优值。

表 4 各层添加辅助分类器作用对比

Tab.4 Comparative results of adding auxiliary classifiers at each layer

Number	Convolution layer of auxiliary classifier	Accuracy
(1)	No auxiliary classifier	0.800
(2)	Layer (3a)	0.817
(3)	Layer (3b)	0.808
(4)	Layer (4c)	0.798
(5)	Layer (4e)	0.815

通过消融实验进一步证实文中结合可视化与人类情感认知添加的辅助分类器对情感分类的贡献。通过删除 Layer (4a)、Layer (4b)、Layer (4d) 层的辅助分类器，并保持其他结构不变，构建实验所需的 6 种消融网络结构。实验结果如表 5 所示。消融实验 (2)~

表 5 消融实验结果对比

Tab.5 Ablation experiment results

Number	Ablation network structure	Accuracy
(1)	Our algorithm	**0.881**
(2)	Delete the auxiliary classifier of layer (4a)	0.862
(3)	Delete the auxiliary classifier of Layer (4b)	0.866
(4)	Delete the auxiliary classifier of Layer (4d)	0.858
(5)	Delete the auxiliary classifier of Layer (4a) and Layer (4b)	0.849
(6)	Delete the auxiliary classifier of Layer (4a) and Layer (4d)	0.845
(7)	Delete the auxiliary classifier of Layer (4b) and Layer (4d)	0.851

(4) 分别删除文中算法中的 1 个辅助分类器，与未删除算法相比准确率均有下降。实验 (5)~(7) 分别删除 2 个辅助分类器，显然准确率下降更为严重。说明正如特征可视化所示，文中选择的 3 层卷积层含有大量情感语义信息，为其添加辅助分类器能够有效控制网络重点学习方向，所构建的模型在自然图像情感识别任务中具有优势。

3.2 可视化实验与分析

仅根据实验数据结果对模型进行定量分析具有局限性。文中将训练完成的情感分类模型“解剖”显示，把网络学习到的自然图像情感特征可视化为人类能够理解的图像，以便分析模型分类的原因，进一步在人类感知层面上证明其情感分类能力。

3.2.1 类激活映射可视化实验与分析

文中使用梯度加权类激活技术[11]对影响输出类别的输入图像的像素进行突出显示，探究图像在分类过程中不同位置对结果的影响大小，判断模型提取特征的来源是否符合预期。如果类激活图显示人类主观认为能够表现图像情感的区域激活较大，即网络确实学习了自然图像中的情感相关信息，那么分类结果真实可信。若类激活图显示网络对无关图像情感的区域更感兴趣，例如图片水印等，则说明网络并未针对情感训练数据，而是由于数据集具有某种尚未发现的特点，导致网络学习非情感相关的特征也能获得较高的准确率。综上所述，类激活图分析能够有效防止 CNN 利用较好的实验结果数据“欺骗”研究者。

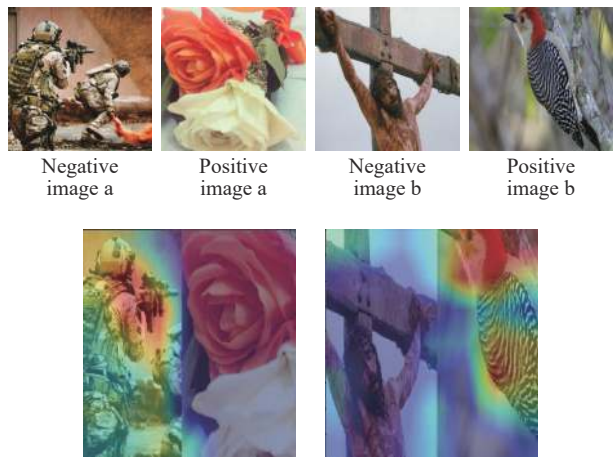
图 5 为一幅正类示例图像分别在网络 Layer (4b) 层和 Layer (5b) 层的类激活图，展示了在这两层卷积层中，图像的不同位置对分类结果的影响大小。图中将原图像和类激活映射结果叠加显示，红色表示该区域对图像分为正类的结果贡献最大，蓝色区域对分类结果贡献小。图 5(a) 显示在网络检测中前期的 Layer (4b) 层，图像中微笑的嘴唇、眼睛、金发等区域对分类贡献大，显然该层在人判断情感时通常更关心的对象位置激活更大，符合人类认知规律。图 5(b) 最深的 Layer (5b) 层类激活图显示大部分区域都对分类有影响，说明是图像的整体情感氛围决定了最终分类节点的结果。可视化结果证明，文中算法成功利用人类的情感知识优化了网络，充分利用中间层不仅有效地弥补了深层网络丢失的信息，而且强化了图像展现情感区域的引导分类作用，帮助提高网络识别情感因素的能力。

(a) Layer (4b) (b) Layer (5b)

图 5 正类示例图像类激活图

Fig.5 Class activation map of a positive class example image

通过正、负类图拼接而成的图像进一步证明模型识别图像情感的能力。图 6 为两幅拼接图像在 Layer (5b) 层的类激活图。可视化实验输入图像均由一幅负类图像 (左侧) 和一幅正类图像 (右侧) 拼接而成。图 6(a) 被分为负类，可以观察到图像中左侧的负类图像区域对分类结果做出贡献，而右侧部分基本对结果无影响。同样，图 6(b) 中右侧区域对于图像被分为正类做出了更大贡献。类激活图显示文中改进的模型的确识别了图像的情感因素，网络运作行为符合人类预期。



(a) 拼接图像 a (b) 拼接图像 b
(a) Spliced image a (b) Spliced image b

图 6 拼接图像类激活图

Fig.6 Class activation map of spliced images

3.2.2 空间位置可视化实验与分析

接下来文中对各空间位置的神经元组合进行可视化，解释网络分类原因。图 7 显示负类图像 7(a) 分别在 Layer (4b)、Layer (4d)、Layer (5b) 层的空间位置特征可视化。整体特征由图片各位置的特征单元格

组成，单元格数量取决于相应卷积层的尺寸，Layer (4b)、Layer (4d) 层的空间位置特征图 7(b)、7(c) 由 14×14 个单元格组成，Layer (5b) 层的图 7(d) 由 7×7 个单元格组成。观察可知在 Layer (4b) 层 (如图 7(b)) 不同物体对象可视化效果不同，例如人物区域和背景区域特征有明显差别。说明网络检测中前期主要根据物体对象的形状、颜色等方面提取情感特征，此时物体的特点及其情感象征意义主导情感分类。图 7(c)、7(d) 显示网络越深，图像不同位置的特征趋近相似，说明网络深层提取了图像整体氛围的高级特征，据此判断图像情感。

(a) Original image (b) Layer (4b)

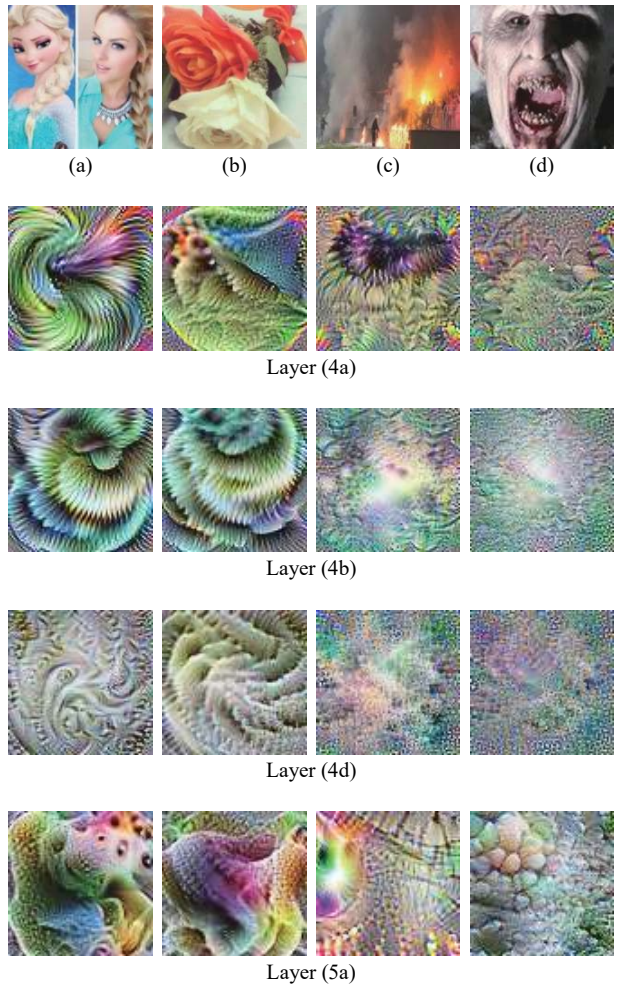
(c) Layer (4d) (d) Layer (5b)

图 7 负类示例图像空间位置特征可视化

Fig.7 Visualization of spatial location features of a negative class image

3.2.3 神经元组可视化实验与分析

文中通过可视化神经元组将网络各层学习到的情感特征可视化为人能够理解的图像，使网络运作行为肉眼清晰可辨。图 8 显示正类图像 8(a)、8(b) 和负类图像 8(c)、8(d) 分别在 Layer (4a)、Layer (4b)、Layer (4d)、Layer (5a) 层的神经元组可视化，直观地表现网络各层提取到的情感信息。同类图像虽然描述对象完全不同，但其神经元组特征表现出了惊人的相似性。异类图像的特征差异明显。正类图像神经元组可视化图给人柔和饱满之感，线条流畅，纹路精致有序，色彩过渡和谐；负类图片的神经元组可视化整体观感很“渣”，画面布满琐碎的颗粒，纹理相对错综凌乱，颜色杂乱无章。显然，非负矩阵分解获取到了为情感识别做出突出贡献的神经元组合，其可视化图蕴含强烈的感情语义色彩，易于观察识别。由于网络训练得到了这样的一组与情感因素密切相关的神经元，输入图像使得相应神经元组高激活，引导模型判断图像情感。进一步验证了文中算法图像情感识别的有效性。



(a) (b) (c) (d)

Layer (4a)

Layer (4b)

Layer (4d)

Layer (5a)

图 8 神经元组特征可视化

Fig.8 Visualization of neuron group features

4 结 论

文中提出一种数据驱动和知识驱动相结合的自然图像情感分类算法，通过可视化技术实现人类与分类网络的知识交互，从而优化深度网络结构。一方面 CNN 通道和神经元组特征可视化帮助人类观察了解深度网络提取的情感特征，根据人类主观感知的图像情感语义信息调整网络结构，充分利用中间层情感信息，从而提高自然图像情感识别鲁棒性。另一方面，训练完成后借助梯度加权类激活映射技术、空间位置

和神经元组特征可视化将模型工作过程与分类原因反馈给研究者,直观地验证了网络训练正确遵循人为控制的学习侧重点,进一步证明算法的情感识别能力,增强其可信性和可解释性。

文中研究成果可用于用户个性化推荐、社交舆情监测分析,提高智能计算的决策透明度和行为可预知性,有助于加强对人工智能技术的安全监管,推动人工智能技术更好地服务人类社会。

参考文献:

[1] Yao A, Shao J, Ma N, et al. Capturing au-aware facial features and their latent relations for emotion recognition in the wild[C]//Proceedings of the ACM International Conference on Multimodal Interaction, 2015: 451-458.

[2] Kang D, Shim H, Yoon K. A method for extracting emotion using colors comprise the painting image [J]. ***Multimedia Tools and Applications***, 2018, 77(4): 4985-5002.

[3] Song K, Yao T, Ling Q, et al. Boosting image sentiment analysis with visual attention [J]. ***Neurocomputing***, 2018, 312: 218-228.

[4] He X, Zhang W. Emotion recognition by assisted learning with convolutional neural networks [J]. ***Neurocomputing***, 2018, 291: 187-194.

[5] Castelvecchi D. Can we open the black box of AI? [J]. ***Nature***, 2016, 538(7623): 20-23.

[6] Voosen P. How AI detectives are cracking open the black box of deep learning[EB/OL]. [2019-09-12]. https://doi.org/10.1126/science.aan7059.

[7] Fu R, Li B, Gao Y, et al. Visualizing and analyzing convolution neural networks with gradient information [J]. ***Neurocomputing***, 2018, 293: 12-17.

[8] Olah C, Mordvintsev A, Schubert L. Feature visualization[C]// IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 2034594.

[9] Olah C, Satyanarayan A, Johnson I, et al. The building blocks of interpretability [J]. ***Distill***, 2018: 00010.

[10] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921-2929.

[11] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the International Conference on Computer Vision, 2017: 618-626.

[12] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Proceedings of the European Conference on Computer Vision, 2014: 818-833.

[13] Fong R, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation[C]//Proceedings of the International Conference on Computer Vision, 2017: 3449-3457.

[14] Li Yuzhi, Sheng Jiachuan, Hua Bin. Improved embedded learning for classification of Chinese paintings [J]. ***Journal of Computer-Aided Design & Computer Graphics***, 2018, 30(5): 893-900. (in Chinese)

[15] Sheng Jiachuan, Li Yuzhi. Learning artistic objects for improved classification of Chinese paintings [J]. ***Journal of Image and Graphics***, 2018, 23(8): 1193-1206. (in Chinese)

[16] Liu Pengfei, Zhao Huaici, Cao Feidao. Blind deblurring of noisy and blurry images of multi-scale convolutional neural network [J]. ***Infrared and Laser Engineering***, 2019, 48(4): 0426001. (in Chinese)

[17] Fang Shengnan, Gu Xiaojing, Gu Xingsheng. Infrared target tracking with correlation filter based on adaptive fusion of responses [J]. ***Infrared and Laser Engineering***, 2019, 48(6): 0626003. (in Chinese)

[18] Hu Shanjiang, He Yan, Tao Bangyi, et al. Classification of sea and land waveforms with deep learning for airborne laser bathymetry [J]. ***Infrared and Laser Engineering***, 2019, 48(11): 1113004. (in Chinese)

[19] Sheng J, Li Y. Classification of traditional Chinese paintings using a modified embedding algorithm [J]. ***Journal of Electronic Imaging***, 2019, 28(2): 023013.

[20] Sheng J, Song C, Wang J, et al. Convolutional neural network style transfer towards chinese paintings[C]//IEEE Access, 2019, 7: 163719-163728.

[21] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations, 2015: 1-14.

[23] Erhan D, Bengio Y, Courville A, et al. Visualizing higher-layer features of a deep network[D]. Canada: University of Montreal, 2009.

[24] Lin M, Chen Q, Yan S. Network in network[C]//Proceedings of the International Conference on Learning Representations, 2014: 1-10.

[25] You Q, Luo J, Jin H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2015: 381-388.