基于 ROI-KNN 卷积神经网络的面部表情识别

孙 晓 [1] 潘 汀 [1] 任福继 [1, 2]

摘 要 深度神经网络已经被证明在图像、语音、文本领域具有挖掘数据深层潜在的分布式表达特征的能力. 通过在多个面部情感数据集上训练深度卷积神经网络和深度稀疏校正神经网络两种深度学习模型, 对深度神经网络在面部情感分类领域的应用作了对比评估. 进而, 引入了面部结构先验知识, 结合感兴趣区域 (Region of interest, ROI) 和 K 最近邻算法 (K-nearest neighbors, KNN), 提出一种快速、简易的针对面部表情分类的深度学习训练改进方案 — ROI-KNN, 该训练方案降低了由于面部表情训练数据过少而导致深度神经网络模型泛化能力不佳的问题, 提高了深度学习在面部表情分类中的鲁棒性, 同时, 显著地降低了测试错误率.

关键词 卷积神经网络, 面部情感识别, 模型泛化, 先验知识

引用格式 孙晓, 潘汀, 任福继. 基于 ROI-KNN 卷积神经网络的面部表情识别. 自动化学报, 2016, **42**(6): 883−891

DOI 10.16383/j.aas.2016.c150638

Facial Expression Recognition Using ROI-KNN Deep Convolutional Neural Networks

SUN Xiao[1] PAN Ting[1] REN Fu-Ji[1, 2]

Abstract Deep neural networks have been proved to be able to mine distributed representation of data including image, speech and text. By building two models of deep convolutional neural networks and deep sparse rectifier neural networks on facial expression dataset, we make contrastive evaluations in facial expression recognition system with deep neural networks. Additionally, combining region of interest (ROI) and K-nearest neighbors (KNN), we propose a fast and simple improved method called "ROI-KNN" for facial expression classification, which relieves the poor generalization of deep neural networks due to lacking of data and decreases the testing error rate apparently and generally. The proposed method also improves the robustness of deep learning in facial expression classification.

Key words Convolution neural networks, facial expression recognition, model generalization, prior knowledge

Citation Sun Xiao, Pan Ting, Ren Fu-Ji. Facial expression recognition using ROI-KNN deep convolutional neural networks. *Acta Automatica Sinica*, 2016, **42**(6): 883−891

面部情感识别是情感计算中情感识别的重要研究内容之一. 面部五官的不同移动、变化程度及其组合, 结合人脑中预存的先验知识, 构成生物情感认知系统中最敏捷、有效的识别部分, 面部表情在情感交互中承载了大部分的信息.

对计算机而言, 面部表情识别是一项艰巨的任务. 计算机想要完成面部表情识别任务, 需要大量的训练数据 (标注的面部表情数据) 来降低模型系统的不确定性. 然而, 目前尚未形成面部情感的自然大数据集 (标注的自然条件下的面部表情数据集), 这就意味着, 现有的面部表情识别模型系统中存在着大量不确定性. 尽管在一个数据集的测试集上表现良好, 但当实际应用时, 模型对随机的新数据泛化能力就会变得很差, 鲁棒性很低.

面部情感识别系统通常包括三部分: 面部数据采集 (标注)、特征提取、情感识别等. 面部数据采集包含人脸检测、人脸关键点标记等两大手段. 在获得数据之后, 进而对数据进行特征提取. 可以使用主成分分析 (Principal component analysis, PCA) 等简易的线性变换方法, 也可以使用常见的人工特征方法, 如尺度不变特征变换 (Scale-invariant feature transform, SIFT)、Haar、局部二值模式 (Local bi-

收稿日期 2015-10-12 录用日期 2016-04-01
Manuscript received October 12, 2015; accepted April 1, 2016
国家自然科学基金重点项目 (61432004), 安徽省自然科学基金 (1508085QF119), 模式识别国家重点实验室开放课题 (NLPR201407345), 中国博士后科学基金 (2015M580532), 合肥工业大学 2015 年国家省级大学生创新训练计划项目 (2015cxcys109) 资助
Supported by Key Program of National Natural Foundation Science of China (61432004), the Natural Science Foundation of Anhui Province (1508085QF119), Open Project Program of the National Laboratory of Pattern Recognition (NLPR201407345), China Postdoctoral Science Foundation (2015M580532), and National Training Program of Innovation and Entrepreneurship for HFUT Undergraduates (2015cxcys109)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 合肥工业大学计算机与信息学院 合肥 230009 中国 2. 德岛大学智能信息工学部 德岛 7708500 日本
1. School of Computer and Information, Hefei University of Technology, Hefei 230009, China 2. Department of Information Science and Intelligent Systems, Faculty of Engineering, Tokushima University, Tokushima 7708500, Japan

nary pattern, LBP) 等. 最后, 将提取到的特征数据输入到判别分类器当中, 得到识别结果.

随着深度神经网络的提出, 图像识别领域的 "先提取特征, 后模式识别" 这一框架被打破. Krizhevsky 等[1] 在 ILSVRC-2012 图像识别竞赛中, 利用深度卷积神经网络的自适应特征提取能力, 使得模型的测试成绩远远超过了 SIFT 等具有旋转缩放不变性的人工特征. 最近, 在面部情感识别任务上, Lopes 等[2] 尝试引入了卷积神经网络模型, 将特征提取和判别分类两个步骤统一结合, 在 Extended CohnKanade (CK+)[3] 静态情感数据集上取得了很好的测试结果. 然而, 目前大多数针对面部表情的深度学习模型是在标准数据集上训练并获得较好的结果, 在实际应用中却出现精度急剧下降, 无法重现实验室模型的准确率, 这部分原因在于基于 CK+ 等标准数据集上训练的模型有两个比较明显的缺陷:

1) 其数据都是摄像机通过正规的角度采集, 这与实际系统获得的 Wild 数据有很大的差别, 会导致模型的泛化能力很差. 如图 1 所示. 在实验部分, 通过设计相应实验验证了这点.

图 1 CK+ 与 Wild 数据集样例

Fig. 1 Samples from CK+ and Wild

2) CK+ 数据集有 593 幅面部表情图像, 包括愤怒、厌恶、恐惧、高兴、悲伤、惊讶六种基本情绪, 这意味着平均每种表情有不到 100 个训练样本.

即便引入非表情峰值的图像进行扩展, 或采用 Lopes 等[2] 的对单张图片旋转采样生成 30 张的扩展训练样本方法, 最后得到的大部分图像携带的信息都有重复 (接近于样本简单复制), 与同数量的不同样本在信息量上还是有不少差距. 而目前小型图像数据集的原始数据量基准都是 60 k (不包括采样生成), 如 MNIST、Cifar10 等数据集. 相对这些数据集, 在 CK+ 上的训练更容易达到过拟合.

鉴于以上两点问题分析, 目前基于 CK+ 数据集训练并得到的高准确率 (95 %) 测试结果并不意味着当前模型系统已经胜任真实的面部情感识别任务, 或超越人类的识别结果. 本文第 1 节将介绍深度神经网络在模型结构上的一些新变化. 第 2 节将介绍两种基本的深度神经网络结构以及针对小数据集的先验改良方法. 第 3 节介绍混合 CK+ 与从互联网上收集 Wild 面部表情数据形成的新数据集, 包含实验测试结果与分析. 第 4 节是结论、归纳. 本文中使用的基于 Theano 深度学习框架的相关代码和训练参数可从 Github[1] 获取.

1 相关工作

1.1 神经网络

神经网络的出发点是 "参数" 拟合 "函数", Bishop[4] 从贝叶斯概率体系角度证明了拟合学习算法的判别根据:

$$p(t|x,t',\alpha,\beta)=\mathrm{N}(t|m_N^{\mathrm{T}}\Phi(x),\sigma_N^2(x)) \quad (1)$$

$$m_N^{\mathrm{T}}\Phi(x)=y(x,m_N)=\sum_{n=1}^{N}kernel(x,x_n)t_n \quad (2)$$

式 (1) 表明了预测数据 t, 在训练数据 t'、x, 以及训练数据高斯方差 β、参数高斯方差 α 的概率分布同样是一个高斯分布. 式 (2) 表明了该高斯分布的均值为一个等价核函数 (即 Smooth 矩阵) 与训练目标的乘积. 该核函数衡量着预测输入 x 与训练输入 x_n 的距离. 距离越近, 数值越大, 预测目标 t 就越接近训练目标 t_n, 反之亦然.

Bengio[5] 指出, 参数模型如支持向量机 (Support vector machine, SVM)、浅层神经网络, 非参数模型如 K 最近邻算法 (K-nearest neighbors, KNN), 最基本的特性都是基于训练样本与预测样本输入的空间距离而做出预测结果的, 称之为平滑先验 (Smoothness-prior). 这个先验在目标函数随输入空间变化敏感时, 只能采集到局部特征 (Local representation), 会得到很差的泛化结果, 而图像数据的输入空间恰好如此. 因而, 不可以直接在图像任务中使用这些分类器, 而需要先提取特征. 从流形学习的观点看, SIFT、Haar、LBP 等人工特征或是 PCA 这类的简单线性变换特征将输入空间的流形面从高维降至低维, 如图 2, 由于流形面是局部光滑的, 从而使得具有平滑判别能力的分类器在流形面区域变换后, 仍然可以很好地分类.

1.2 深度卷积神经网络

LeCun 等[6] 在 1990 年提出的深度卷积神经网络, 如图 3. 以 Fukushima[7] 的感知机结构为基础, 借助 Rumelhart 等[8] 的反向传播训练算法, 首先在文字图像识别领域取得巨大成功[9].

卷积神经网络与一般的全连接式神经网络相比较, 除了在模型中注入 Smooth 这样的先验知识之外, 还注入一些针对图像数据特点的先验知识.

[1] https://github.com/neopenx/

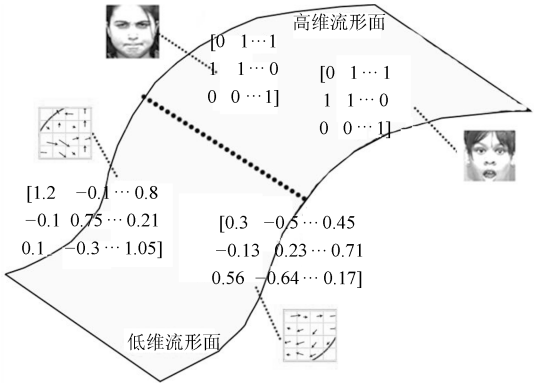

图 2　输入空间的流形面

Fig. 2　Manifold side of input space

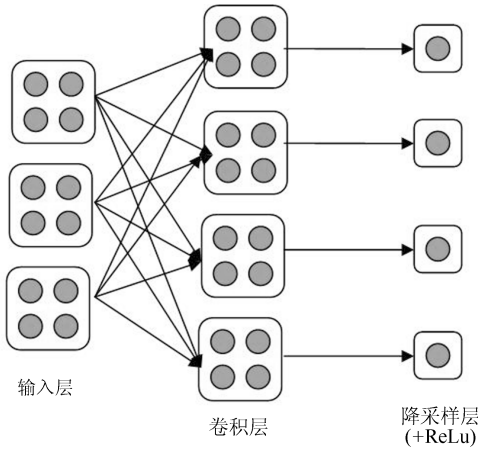

图 3　卷积神经网络的局部块状连接与基本结构

Fig. 3　Local connection and structure of convolutional neural network (CNN)

1.2.1　局部性

图像中包含的隐含信息在输入空间内具有局部平滑性, 因此卷积神经网络针对像素块建立块状神经元的局部连接 (Locally-connection). 传统基于像素点的连接称为全连接 (Fully-connection) 或稠密连接 (Dense-connection). 块状神经元显著减少了每层神经元参数个数, 这使得误差从输出层开始, 以较小的广度发散[5], 同时可以增加神经网络深度, 来保持结构中深度和广度之间的平衡. Szegedy 等[10]利用此特性构建出 22 层的 GoogLeNet, 赢得了 ILSVRC-2014 图像识别竞赛冠军.

1.2.2　权值共享/局部感受野

当二维神经元块维度小于二维数据块 (特征图) 时, 意味着神经元块参数会在数据块的不同区域重复加权计算, 这构成权值共享, 数学形式即二维离散卷积. 权值共享的做法借鉴了视觉神经感受野的概念, Fukushima[7] 认为局部感受野使得模型获得图像中的平移不变性, 增强泛化能力.

1.2.3　降采样

降采样 Pooling 层是一个非参数层, 作用是将一定邻域内的像素块压缩成一个像素点, 使图像缩放. 它通常紧接着卷积层, 根据缩放算法的不同, 分为锐化 (Max pooling)、平滑 (Avg pooling). 对输入图像数据块的逐层缩放, 让各层获得不同比例的局部感受野, 使得模型获得图像中的缩放不变性, 增强泛化能力.

1.3　深度稀疏校正神经网络

Glorot 等[11] 提出深度稀疏校正神经网络 (Deep sparse rectifier neural networks) 从结构上仍然属于全连接神经网络, 唯一变化是将 Sigmoid 型 (logistic/tanh) 激活函数全部替换成了 ReLU.

1.3.1　深度结构的有效性

Barron[12] 证明了拥有一个隐层、N 个神经元的全连接神经网络可以将任何函数拟合至 $1/N$ 精度. 这意味着, 如果需要增加拟合精度, 只要广度, 而无需考虑深度. 而 Bengio[5] 认为如果一个函数可以由多个函数组合得到, 在数据有限的情况下, 使用过浅的深度会影响拟合的效果, 引起训练周期过长、泛化能力很差等问题. Hubel 等[13] 在实验中发现猫的视觉皮层由多层抽象神经结构完成, V1 层提取图像边缘特征, V2 层开始逐层组合出部分形状, 直至最后组合抽象出完整的视觉目标. 这从生理学角度证明了图像识别函数可以由多个函数组合而成, 增加神经网络的深度要比广度有效得多.

1.3.2　ReLU 激活函数

Dayan 等[14] 通过拟合数据, 发现生物神经元输入电信号与激活频率之间的函数图像具有相对的不对称性与对称性, 如图 4 所示, 不对称区域出现了一段很突兀的 “0”, 这与主流的 Sigmoid 函数有很大区别, 而与 ReLU 函数比较相似. Attwell 等[15]通过实验观察到, 脑神经元在一定时刻, 平均只有 1%～4% 被激活, 这段 “0” 起到了很强的校正作用, 让大部分神经元处于完全不激活状态, 这是生物神经网络具有数以千亿计的神经元, 而不会像模型神经网络一样引发 “维数灾难” 的原因. ReLU 激活函数定义为:

$$ReLU(x) = \max(0, x)$$

Softplus 函数是它的平滑版本:

$$Softplus(x) = \log(1 + \mathrm{e}^x)$$

Softplus 与 ReLU 都是非饱和函数, 它们输出的上下界不被限制在 $[-1, 1]$ 之内, 这大大缓解了深度结构带来的梯度发散 (Gradient vanish) 问题, 促进梯

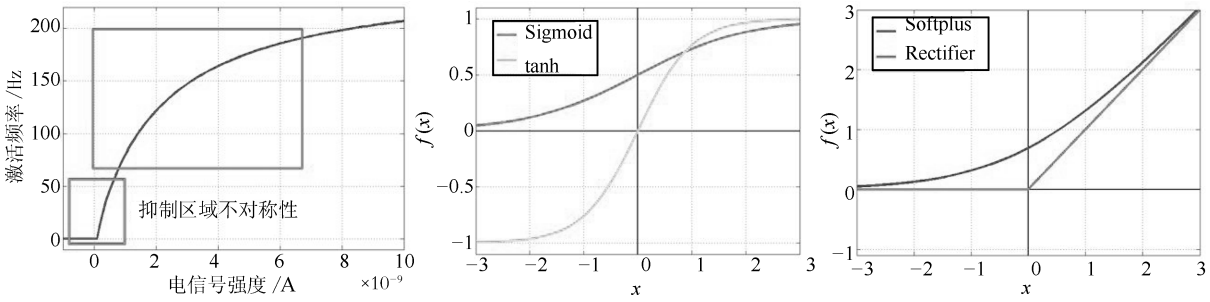

图 4 不同激活函数的函数图像 (图片源自 Glorot[11])

Fig. 4 Graphs for different activation functions from Glorot[11]

度在反向传播中路径中的流动, 在训练庞大神经网络时, 有数倍的提速效果[1]. 另外, 校正 "0" 为模型注入了大量稀疏性, 与 L1 Regularization 效果相同. 目前已知, 稀疏性有助于推动线性不可分转化为线性可分, 缩小做逐层贪心预训练与不做之间在泛化效果上的差异[11].

1.4 Dropout

Hinton 等[16] 提出的 Dropout 层在大量实验中已经被证实可以有效改善任意神经网络的过拟合问题. Dropout 分为两个阶段:

1) 训练阶段: 此时经过该层的所有输入点 x, 都会以一定概率 p 被置为 0, 即该神经元被剔除. 定义式:

$$DropoutTrain(x) = RandomZero(p) \times x$$

这是一个随机过程, 意味着每次正向传播, 网络的有效结构都会产生变化.

2) 测试阶段: 此时应该激活所有神经元, 变成完整结构. 激活全部神经元等效于多个随机神经网络的叠加求和, 需要对输入 x 做一个平均处理, 不然会出现数值问题. 定义式:

$$DropoutTest(x) = (1-p) \times x$$

Dropout 能有效改善过拟合可以从两个角度理解. 首先, Dropout 引入了随机化的稀疏性, 让庞大的神经网络模型在同一时刻只工作一部分, 这与 Attwell 等[15] 在生物神经方面的工作不谋而合. 其次, 由于每次网络的结构都在变化, 参数会不停受到惩罚, 被迫向一个稳定的方向调整, 而不是简单地做拟合. 这与 Darwin[17] 提出的 "自然竞争选择" 概念切合, 拉近了模型神经网络与生物神经网络的距离.

1.5 初始化

1.5.1 权值初始化

传统的神经网络权值初始化为:

$$W = Uniform\left(-\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}\right)$$

Xavier 等[18] 提出了更适合 Sigmoid 函数的方案:

$$W = Uniform\left(-\frac{1}{\sqrt{Fin+Fout}}, \frac{1}{\sqrt{Fin+Fout}}\right)$$

其中, Fin 为输入维度, $Fout$ 为输出维度. Bishop[4] 指出, 在 $N \to \infty$ 时, 均匀分布会演变为高斯分布, 更一般地, 任意连续的随机变量都可以假定符合高斯分布. 而贝叶斯拟合模型引入的关于 W 的共轭先验分布也是假定 $P(W)$ 服从高斯分布. 这意味着, 使用均匀分布来初始化 W 不是一个很好的方案. Krizhevsky 等[1]、Hinton 等[16] 在 ILSVRC-2012 图像识别竞赛的冠军模型中, 对 W 的初始化使用了零均值、常数方差的高斯分布而不是传统的均匀分布, 从实验角度证明了高斯分布初始化的合理性.

1.5.2 偏置初始化

Krizhevsky 等[1]、Hinton 等[16] 将神经网络隐层 (非输出层) 的偏置初始化为 1 而不是 0, 让训练在初期阶段得到很大加速. 目前尚无数学解释, 只是经验规则.

2 结构、超参数与改良方法

2.1 深度卷积神经网络

如图 5, 针对输入大小为 32 × 32 的灰度图 (彩色维度为 1), 构建了 3 个卷积与 Max pooling 层、1 个全连接层、1 个 Softmax 层. 根据各层神经元个数的不同, 又分为 CNN-64、CNN-96、CNN-128.

CNN-64: $[32, 32, 64, 64]$

CNN-96: $[48, 48, 96, 200]$

CNN-128: $[64, 64, 128, 300]$

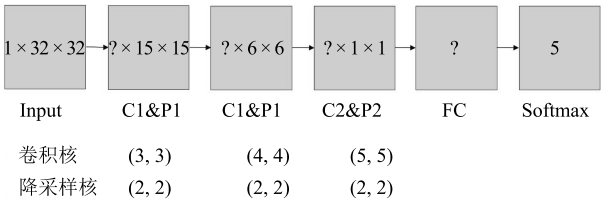

图 5　深度卷积神经网络的结构
(? 表示不确定超参数, 有多种优选方案.)

Fig. 5　Structure of DNN
(? represents uncertain parameters with many candidate solutions.)

为了减轻过拟合问题, 全连接层后连接着一个 $p=0.5$ 的 Dropout 层, 而不是使用 L2 Regularization.

除 Softmax 层之外, 其余各层激活函数均为 ReLU, 卷积层输出激活后, 再输入到 Max pooling 层. 权值 W 的初始化采用 Krizhevsky 等[1] 的零均值、常数标准差 (Standard deviation, STD) 方案. 各层 STD 分别为:

[0.0001, 0.001, 0.001, 0.01, 0.1]

偏置的初始化采用 Krizhevsky 等[1] 的方案.

2.2　深度稀疏校正神经网络

如图 6, 针对输入大小为 32 × 32 的灰度图 (彩色维度为 1), 构建了 3 个全连接层、1 个 Softmax 层.

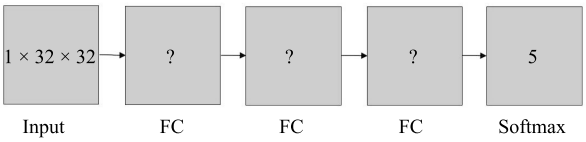

图 6　深度稀疏校正网络的结构

Fig. 6　Structure of deep sparse rectifier net

根据各层神经元个数的不同, 又分为 DNN-1000, DNN-2000.

DNN-1000: [1 000, 1 000, 1 000]

DNN-2000: [2 000, 2 000, 2 000]

为了减轻过拟合问题, 三个全连接层后各连接着三个 $p=0.2$ 的 Dropout 层. 除 Softmax 层之外, 其余各层激活函数均为 ReLU. 权值 W 的初始化各层 STD 分别为: [0.1, 0.1, 0.1, 0.1].

在测试中发现, 隐层偏置全部设为 1 对于深度稀疏校正神经网络效果并不好, 所以设为 0.

2.3　数据预处理、训练参数控制

本文的数据处理中只做了均值标准化, 取训练数据 32 × 32 的各个维度计算 1 024 个均值并序列化保存. 训练、测试时, 减去均值. 特别地, DNN 在均值标准化后, 对数值缩小 128.0 倍.

训练过程中使用交叉验证与早终止 (Early stopping). 两个模型学习率 lr 为常数 0.01, 动量 momentum 为常数 0.9. 交叉验证中发现验证集错误率不再下降或上升时, 即判定为学习率 lr 过大, 停止并降低一个数量级, 再次训练, 重复直至学习率在 0.0001 阶段结束, 经历 3 个数量级的训练.

2.4　ROI-KNN

Xavier 等[18−19] 在利用深度卷积神经网络训练人脸特征时, 采取对单张图片不同尺度区域切割的方法, 来扩大数据集. 本文借鉴了此方法, 并针对面部表情识别做了改进, 根据人脸的面部结构, 设置了 9 个不同的感兴趣区域 (Region of interest, ROI), 如图 7, 主动引导神经网络关注与表情相关的面部区域.

图 7　9 个 ROI 区域 (切割、翻转、遮盖、中心聚焦)

Fig. 7　Nine ROI regions (cut, flip, cover, center focus)

设置 ROI 区域使用的都是图像处理中的基本手段, 包含切割、翻转、遮盖、中心聚焦. 为了确保不同面部的 ROI 区域不会有太大偏差, 需预先进行人脸检测提取人脸, 使面部填充图像的大部分区域, 让面部中轴线近似与图像中轴线重合.

切割方案重点关注眼、鼻、嘴在不同表情中的区别, 为了尽量让处理手段简单, 并没有预先检测面部关键点来切割. 翻转方案考虑了拍摄方式的不同. 遮盖方案是对切割方案的联合. 中心聚焦方案去除了一定噪声 (如头发).

ROI 方法让训练数据扩大至 9 倍, 这种扩大是否有效, 取决于这些迥异 ROI 区域之间是否存在着某些联系, 有助于增强预测目标的信度. 这里的增强更强调 ROI 区域对测试原始图像的增强、不同 ROI 区域之间的增强 (如左眼对上半脸), 而不仅仅是相同 ROI 区域间的增强 (如左眼对左眼). Bengio[5] 指出了这两者的区别: 前者的成功源于模型挖掘出了分布式表达特征 (Distributed representation), 分布式表达特征让模型对未观测的数据有着很好的泛化和归纳. 而后者的成功则受 Smooth-prior 作用下的局部表达特征 (Local representation) 影响较大, 与训练数据、测试数据在输入空间的距离有很大关

系. 在下一节的实验会证明 ROI 区域确实对判别原始图像有一定增强.

ROI 数据倍增的效果是针对训练阶段的, 而在测试阶段最直接的方法是对测试图像直接判别. 但因为这会浪费模型中记忆的关于 ROI 区域的分布式表达特征. 尽管这些特征在直接做判别时具有推动作用, 但未起到更大作用.

众多机器学习模型中, KNN 具有出色且简单的归并能力, 它通过建立贪心投票机制, 让多个判别目标联合, 缩小最终的判别范围, 强化最终的判别信度. 鉴于此, 提出 ROI-KNN 方法, 在测试时, 对 9 个 ROI 区域的判别结果投票, 取票数最多的判别结果作为最终结果, 在线归并原始结果.

ROI-KNN 的最大缺陷是对原始模型训练的 Distributed representation 有很高的要求, 因为这些 ROI 的输入信息较完整图片要小很多, 直观上来看, 就是放大关注细节. 训练 ROI 与测试 ROI 之间细微的差别, 被放大之后, 模型中的 Local representation 会对判别产生很大干扰. 在实验中最直接的体现就是 ROI 区域的测试错误率要大于原始图像错误率, 若基于这种情况下投票, 那么最后的投票结果反而比不投票要差. 下一节将设计相应的实验验证.

2.5 旋转生成采样

Lopes 等[2] 扩大数据集的方法是将原始图像轻微旋转一定角度, 生成大量变化的训练样本. 这种做法看起来似乎是没有问题的, 因为深度卷积神经网络本身具有挖掘图像缩放不变性、平移不变性的能力, 唯独缺少旋转不变性.

在这里必须考虑一个问题: 强行注入旋转变换的样本能否让模型获得旋转不变性? 本文对此的答案是否定的. 卷积神经网络得到的平移、缩放不变性是模型不断提炼、泛化的成果, 而直接注入的旋转样本可能只会让模型出现过拟合, 因为模型本身并没有提炼旋转不变性的能力, 而本文提出的 ROI 方法是基于平移、缩放不变性的, 没有这种潜在问题. 如果测试数据与训练数据较为接近, 那么过拟合问题就不会暴露, 反之亦然. 本文认为 Lopes 等[2] 注入旋转样本后的模型有过拟合的可能, 因为他们的测试数据与训练数据很接近, 注入旋转样本得到的改善很有可能是过拟合得到的. 在下一节的实验中会使用旋转样本, 对 Wild 数据进行测试来验证.

3 实验

本节使用第 2.1 节和第 2.2 节中构建的两个深度神经网络模型做对比评估, 评估环节的目标包括: ROI 辅助评估、旋转生成样本评估和 ROI-KNN 辅助评估. 最后将评估深度学习模型与非深度学习模型.

3.1 数据集

为了解决 CK+ 数据集过于正规的问题. 从互联网各大搜索引擎中收集了 4 类, 每类 500 张 Wild 数据, 分别是高兴、悲伤、惊讶、愤怒. 此外, 由于 CK+ 数据集的原始类别标签不含有 “中性” 表情, 从合肥工业大学教务管理系统中抽取了 1 200 张学生面部照片, 这些照片除了表情呈中性之外, 与 CK+ 一样, 都是很正规的摄像机取景, 方便在测试集中与 Wild 数据作对比评估. 训练集由 CK+ 的高兴、悲伤、惊讶、愤怒各 700 张混合互联网下载的图片各 200 张以及 “中性” 的 900 张构成. 共计 5 类, 每类 900 张图片. 测试集由互联网下载的图片各 300 张混合 “中性” 的 300 张构成. 共计 5 类, 每类 300 张图片.

3.2 ROI 辅助评估

ROI 辅助评估是本文关注的重点, 它反映着模型内部 Distributed representation 的训练情况. 使用的是第 3.1 节给出的 5 类共 4 500 张面部训练数据、5 类共 1 500 张测试数据. 训练 4 500 张数据经过 ROI 处理后, 为 $4\,500 \times 9 = 40\,500$ 张, 测试数据不做变化. 实验结果如表 1, 基准为无 ROI 强化, “∗” 表示 ROI 强化. 从整体实验结果来看, ROI 的引入对两套模型的各个规模都有 4 %～5 % 的精度提升, 符合预期. 深度卷积神经网络随着规模的提升, 效果也在提升, 达到最好的整体错误率 25.8 %. 逐一对各个表情分析, 可以发现一些问题. 首先, 就是中性测试集相对于其他测试集, 测试成绩非常高. 这是在第 3.1 节数据有意如此设置: 测试集里, 只有中性集没有使用 Wild 数据, 而选择了与训练集较为相似的正规数据, 这个成绩符合预期, 同时证明了 Lopes 等[2] 基于 CK+ 的高准确率测试结果并不一定意味着模型拥有良好的泛化能力. 其次, 悲伤测试集表现最差, 这与 Lopes 等[2] 的结果一致, 说明面部悲伤情感比较难被准确识别, 而高兴、惊讶、愤怒的测试结果则比较接近.

表 1 ROI 辅助评估的测试集错误率 (%)

Table 1 Test set error rate of ROI auxiliary (%)

	中性	高兴	悲伤	惊讶	愤怒	整体
CNN-64	4.7	32.7	54.3	33	40.3	33.3
CNN-64*	5.6	36.3	59.3	20.0	31.7	30.6
CNN-96*	5.0	36.7	53.3	20.7	24.7	28.6
CNN-128	3.3	32.0	51.0	27.0	37.7	30.2
CNN-128*	3.0	31.0	55.7	18.7	24.3	26.6
DNN-1000	3.0	37.7	65.3	38.3	36.7	36.2
DNN-1000*	2.3	39.0	52.0	30.0	31.7	31.0
DNN-2000*	2.0	43.3	55.0	24.7	32.7	31.5

3.3 旋转生成样本评估

在第 2.5 节推测旋转采样生成的样本可能会导致神经网络模型产生过拟合, 为了验证该假设的可能性, 设计了两份新的训练数据:

1) 数据集 I. 针对 CK+ 与高考录取照片两类正规数据, 以图像中心为原点, 进行旋转采样. 旋转方法同文献 [2], 令旋转角 α 服从零均值高斯分布: $\alpha \sim \mathrm{N}(0, 3^{\circ})$ 对源训练集 5 类, 每类 700 张执行高斯随机数 11 次, 加上第 3.1 节 4 500 张训练图像, 共有 $5 \times 700 \times 11 + 4\,500 = 43\,000$ 张, 构成新训练集, 测试集不变化.

2) 数据集 II. 将数据集 I 中的 43 000 张采样数据, 与第 3.2 节中的 40 500 张数据混合, 共计 83 500 张训练数据, 构成新训练集, 测试集不变化. 以第 3.2 节中的无 ROI 测试结果作为对比基准, 实验结果如表 2, "*" 表示使用数据集 I, "+" 表示使用数据集 II+ROI, "∧" 表示使用数据集 II 结合 ROI-KNN.

表 2　旋转生成样本评估的测试集错误率 (%)

Table 2　Test set error rate of rotating generated sample (%)

	中性	高兴	悲伤	惊讶	愤怒	整体
CNN-128	3.3	32.0	51.0	27.0	37.7	30.2
CNN-128*	4.7	41.3	52.7	32.7	35.0	33.2
CNN-128+	3.0	37.0	51.7	15.7	24.0	26.3
CNN-128ˆ	0.0	30.0	54.0	13.0	26.7	24.7
DNN-1000	3.0	37.7	65.3	38.3	36.7	36.2
DNN-1000*	1.3	39.7	62.0	37.3	42.0	36.5
DNN-1000+	2.3	41.3	57.0	30.0	35.7	33.3
DNN-1000ˆ	1.3	43.0	67.7	31.0	33.7	35.3

从整体实验结果来看, 旋转生成样本的引入暴露了不少问题. 首先, 对于数据集 I, CNN-128、DNN-1000 用 43 000 张原始与生成的混合大数据, 得出了比 4 500 的小数据还差的结果, 说明 38 500 张旋转生成样本不仅没有促进归纳和泛化, 反而对 Wild 数据的直接判别产生了干扰, 这与 Lopes 等[2] 的结果截然相反, 本文认为是基于 CK+ 的测试集掩盖了过拟合问题. 其次, 对于数据集 II, ROI 的引入几乎抵消了旋转样本的影响, 但是此时 ROI-KNN 的效果不佳, 在 DNN-1000 中尤为明显. 第 3.4 节中的实验结果表明, ROI-KNN 对模型中的 Distributed representation 有很高的要求, ROI-KNN 的效果不佳, 从另一个角度表明了引入旋转生成样本可能对 Distributed representation 产生了影响. 基于以上两个数据集的测试, 可以判断在面部情感分析任务上, 引入旋转生成样本来扩大数据集并不是一个可取的方案. 它并不能让具有缩放、平移不变性的深度卷积神经网络获得旋转不变性, 反而因为旋转输入空间的引入, 对缩放、平移不变性的效果产生干扰, 构成由于模型挖掘数据能力不足, 导致的不可避免型过拟合, 这种过拟合不是由于参数空间过大引起的, 没有方法通过扩大数据集避免. 当测试数据与训练数据有较大偏差和变化时, 便会显现出来, 若模型训练按照这种方式训练, 则是无法在实际中应用的.

3.4 ROI-KNN 辅助评估

ROI-KNN 辅助评估将考察 KNN 的贪心投票机制对结果的影响, 按照第 2.4 节中的推测, 它对模型内部的 Distributed representation 有很高的要求. 实验结果如表 3, 基准为 ROI 强化, "*" 表示 ROI-KNN 强化.

表 3　ROI-KNN 辅助评估的测试集错误率 (%)

Table 3　Test set error rate with ROI-KNN (%)

	中性	高兴	悲伤	惊讶	愤怒	整体
CNN-64	5.6	36.3	59.3	20.0	31.7	30.6
CNN-64*	1.0	29.7	56.0	17.0	30.0	26.7
CNN-96	5.0	36.7	53.3	20.7	24.7	28.6
CNN-96*	0.3	26.0	56.3	16.0	26.7	25.8
CNN-128	3.0	31.0	55.7	18.7	24.3	26.6
CNN-128*	0.6	22.7	57.0	12.0	26.3	23.7
DNN-1000	2.3	39.0	52.0	30.0	31.7	31.0
DNN-1000*	0.3	37.3	61.0	31.7	31.0	32.2
DNN-2000	2.0	43.3	55.0	24.7	32.7	31.5
DNN-2000*	0.3	40.0	68.0	26.3	33.3	33.6

从整体实验结果来看, KNN 的投票机制让深度卷积神经网络各个规模又得到了 4 %～5 % 的精度提升, 但在深度稀疏校正神经网络中, 不仅没有提升, 反而让整体结果略微变差. 逐一对各个表情分析, 在深度卷积神经网络中, 除了悲伤集外, 其他测试集均有一定提升. 在深度稀疏校正神经网络中, 中性、高兴集有一定提升, 悲伤集变差幅度最大, 其他测试集几乎无变化.

此实验结果表明了 KNN 投票机制对模型的泛化能力 (或 Distributed representation) 有很高的要求, 直接体现在泛化最差的悲伤集上, 各个模型表现均不好. 另一方面, 卷积神经网络整体又比深度稀疏校正神经网络好得多, 可能是得益于内部针对图像处理的先验知识.

3.5 与非深度学习模型的对比

为了比较所提出的 ROI-KNN 方法与 SVM 等非深度学习方法的性能, 设计了另一组实验, 在公开 JAFFE 数据集上, 与 SVM、PCA 等非深度学习方法进行了比较, 其中本文的模型选取了 CNN-128 结合 ROI-KNN. 从表 4 中可以看出, 相对 SVM 等浅层机器学习模型, 本文提出的深度学习模型在传统的数据集上有非常优异的表现.

表 4 在 JAFFE 上的模型对比

Table 4 Comparisons on JAFFE

来源	模型方法	错误率 (%)
Kumbhar 等[20]	Image feature	30 ∼ 40
Lekshmi 等[21]	SVM	13.1
Zhao 等[22]	PCA and NMF	6.28
Zhi 等[23]	2D-DLPP	4.09
Lee 等[24]	RDA	3.3
本文	ROI-KNN+CNN	2.81

4 结论

深度神经网络在面部情感分析任务上具有很大的探索空间. 首先, 在面部情感数据尚未形成大数据集的当下, 如何利用少量的原始数据有效地扩大数据集是一个难题. 本文工作证明了在 Wild 数据测试环境下, 基于 ROI 的数据集扩大策略要比旋转生成扩大策略有效得多; 其次, 现有的神经网络结构在面部表情识别任务上相对人脸识别等其他任务还有很大的上升空间, 如: 在面部表情识别中, 深度卷积神经网络如何获得旋转不变性. 如果模型结构没有泛化数据中某些特性的能力, 同样会造成过拟合, 但不等同于因为参数空间过大而造成的过拟合. 后者可以直接通过在参数上施加惩罚拟合敏感性的 Regularization 解决, 如 L2 Regularization 或 Dropout. 而前者则需要一些先验知识来引导参数朝泛化方向搜索, 如卷积神经网络的局部连接、平移缩放不变性, 或是深度学习的中心思想 "参数逐层贪心预训练初始化", 甚至是循环递归神经网络 (Recurrent neural network, RNN) 中的时序信息. Distributed representation 可能是连接生物神经网络与模型神经网络之间的桥梁, 因为人类的大脑可能没有使用像 SIFT 这样的特征, 而更可能是一种感知整体与整体、部分与部分、整体与部分之间的联系、归纳、泛化的特征. 使用深度神经网络, 虽然可以不使用 SIFT、Haar、LBP 等人工特征, 使用更接近自然特征, 但在处理其内部不可见、不可控、易受影响的 Distributed representation 上, 则需要引入更多的先验知识与处理技巧. 本文提出的 ROI-KNN 方法, 以简易的方式, 间接地利用并观测了模型 Distributed representation 的情况, 对于深度卷积神经网络这样的模型, 有很好的提升效果. 此外, 深度稀疏校正神经网络并非无用武之地, 其计算速度和不俗的精度表现, 仍是硬件条件有限情况的首选.

致谢

本文的实验部分代码实现是基于 Theano[25] 开发的, 在此对其所有的开发和维护者表示感谢.

References

1 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2012. 1097−1105

2 Lopes A T, de Aguiar E, Oliveira-Santos T. A facial expression recognition system using convolutional networks. In: Proceedings of the 28th SIBGRAPI Conference on Graphics, Patterns and Images. Salvador: IEEE, 2015. 273−280

3 Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). San Francisco, CA: IEEE, 2010. 94−101

4 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2007.

5 Bengio Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*. Hanover, MA, USA: Now Publishers Inc., 2009. 1−127

6 LeCun Y, Boser B, Denker J S, Howard R E, Hubbard W, Jackel L D, Henderson D. Handwritten digit recognition with a back-propagation network. In: Proceedings of Advances in Neural Information Processing Systems 2. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990. 396−404

7 Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980, **36**(4): 193−202

8 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

9 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

10 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015. 1−9

11 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, FL, USA, 2011, **15**: 315−323

12 Barron A R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 1993, **39**(3): 930−945

13 Hubel D H, Wiesel T N, LeVay S. Visual-field representation in layer IV C of monkey striate cortex. In: Proceedings of the 4th Annual Meeting, Society for Neuroscience. St. Louis, US, 1974. 264

14 Dayan P, Abott L F. *Theoretical Neuroscience*. Cambridge: MIT Press, 2001.

15 Attwell D, Laughlin S B. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 2001, **21**(10): 1133−1145

16 Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv: 1207.0580, 2012.

17 Darwin C. *On the Origin of Species*. London: John Murray, Albemarle Street, 1859.

18 Xavier G, Yoshua B. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010). Chia Laguna Resort, Sardinia, Italy, 2010, **9**: 249−256

19 Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10 000 classes. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH: IEEE, 2014. 1891−1898

20 Kumbhar M, Jadhav A, Patil M. Facial expression recognition based on image feature. *International Journal of Computer and Communication Engineering*, 2012, **1**(2): 117−119

21 Lekshmi V P, Sasikumar M. Analysis of facial expression using Gabor and SVM. *International Journal of Recent Trends in Engineering*, 2009, **1**(2): 47−50

22 Zhao L H, Zhuang G B, Xu X H. Facial expression recognition based on PCA and NMF. In: Proceedings of the 7th World Congress on Intelligent Control and Automation. Chongqing, China: IEEE, 2008. 6826−6829

23 Zhi R C, Ruan Q Q. Facial expression recognition based on two-dimensional discriminant locality preserving projections. *Neurocomputing*, 2008, **71**(7−9): 1730−1734

24 Lee C C, Huang S S, Shih C Y. Facial affect recognition using regularized discriminant analysis-based algorithms. *EURASIP Journal on Advances in Signal Processing*, 2010, article ID 596842(doi:10.1155/2010/596842)

25 Bastien F, Lamblin P, Pascanu R, Bergstra J, Goodfellow I J, Bergeron A, Bouchard N, Warde-Farley D, Bengio Y. Theano: new features and speed improvements. In: Conference on Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsuper Vised Feature Learning. Lake Tahoe, US, 2012.

孙 晓 合肥工业大学计算机与信息学院情感计算研究所副教授. 主要研究方向为自然语言处理与情感计算, 机器学习与人机交互. 本文通信作者.
E-mail: sunx@hfut.edu.cn
(**SUN Xiao** Associate professor at the Institute of Affective Computing, Hefei University of Technology. His research interest covers natural language processing, affective computing, machine learning and human-machine interaction. Corresponding author of this paper.)

潘 汀 合肥工业大学计算机与信息学院本科生. 主要研究方向为深度学习, 贝叶斯学习理论及其在计算机视觉与自然语言处理方面的应用.
E-mail: neopenx@mail.hfut.edu.cn
(**PAN Ting** Bachelor student at the School of Computer Science and Information, Hefei University of Technology. His research interest covers the theory of deep learning and Bayesian learning, and corresponding applications in computer vision and natural language processing.)

任福继 合肥工业大学计算机与信息学院情感计算研究所教授, 德岛大学教授. 主要研究方向为人工智能, 情感计算, 自然语言处理, 机器学习与人机交互.
E-mail: ren@is.tokushima-u.ac.jp
(**REN Fu-Ji** Professor at the Institute of Affective Computing, Hefei University of Technology and Tokushima University. His research interest coves artificial intelligent, affective computing, natural language processing, machine learning, and human-machine interaction.)