基于兴趣区域深度神经网络的静态面部表情识别

孙 晓,潘 汀
(合肥工业大学计算机与信息学院,安徽合肥 230009)

摘 要: 通过在面部表情数据集上训练深度卷积神经网络、深度稀疏校正神经网络两种模型,对两种深度神经网络在静态面部表情识别方面的应用作了对比和分析.基于面部表情的结构先验知识,提出一种面向面部表情识别的改良方法——K兴趣区域方法,该方法在构建的开放实验数据集上,降低了由于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,使得混合模型普遍且显著地降低了测试错误率.进而,结合实验结果进行了深入分析,并对深度神经网络在任意图像数据集上的可能有效性进行了深入剖析和分析.

关键词: K兴趣区域; 深度神经网络; 深度学习; 面部表情识别
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)05-1189-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.05.023

Static Facial Expression Recognition System Using ROI Deep Neural Networks

SUN Xiao, PAN Ting
(*School of Computer and Information, Hefei University of Technology. Hefei, Anhui* 230009, *China*)

Abstract: By building two models including Deep Convolutional Neural Networks and Deep Sparse Rectifier Neural Networks on facial expression dataset, we made contrastive evaluations in facial expression recognition system with deep neural networks. Based on prior structure knowledge of facial expression, we proposed a fast and simple improved method called K Region Of Interest--'K-ROI', which relieved the poor generalization of deep neural networks on experimental dataset due to insufficient data and decreased the testing error rate apparently and generally. Finally, we infer the experimental results and analyze comprehensively for the possible validity with deep neural networks on arbitrary image dataset.

Key words: K-ROI; deep neural networks; deep learning; facial expression recognition

1 引言

在面部表情识别领域,随着深度神经网络的提出,"先提取特征,后模式识别"的规则被打破. Krizhevsky等人[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,使模型成绩远远超过了SIFT等传统人工特征的成绩.最近,在面部情感识别任务上,Lopes等[2]尝试引入卷积神经网络模型,将特征提取和判别分类两个步骤统一结合,在Extended Cohn Kanade (CK+)[3]静态面部表情数据集上取得了很好的测试结果.然而,基于CK+(如图1所示)数据集训练拍摄角度正规,而且数量少.因而,现有大多数系统基于CK+数据集获得高准确率(95%),并不能断言可以超越人类的识别能力.

本文的第二部分将介绍相关研究工作.第三部分构建两种面向面部表情识别的深度神经网络.第四部分介绍实验测试结果与分析.第五部分对抽象神经网络提出假设计算模型,尝试对实验结果给出合理解释.最后的第五部分是总结与展望.

收稿日期:2016-02-16;修回日期:2016-06-14;责任编辑:蓝红杰
基金项目:安徽省自然科学基金(No.1508085QF119);国家自然科学基金重点项目(No.61432004);模式识别国家重点实验室开放课题(No.NLPR201407345);中国博士后科学基金(No.2015M580532)

图 1 CK+数据与Wild数据对比

2 相关工作

2.1 关注点机制

Jaderberg[4~6]等人提出空间变换网络,可嵌入关注点机制,来引导神经网络学习图像数据中的空间变换不变性. 关注点机制可以通过对训练数据的空间变换生成得到. 他们将空间变换网络用于鸟的检测任务当中. 通过对两组空间变换参数训练,最终验证了关注点学习的有效性:一组参数寻找鸟头,一组参数寻找鸟身,如图 2 所示.

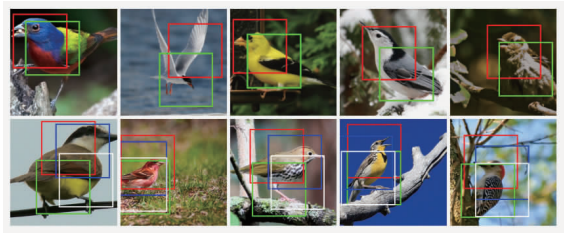

图 2 关注点分别关注鸟的鸟头与鸟身[6]

2.2 深度神经网络

LeCun 等[7]在 1990 年提出深度卷积神经网络,如图 3 所示,以 Fukushima 等[8]的感知机结构为基础,借助 Rumelhart 等[9]的反向传播训练算法,首先在图像识别领域取得巨大成功[10]. 卷积神经网络相比一般的全连接神经网络,除了在模型中注入 Smooth 先验知识之外,还注入一些面向具体任务的先验知识[11].

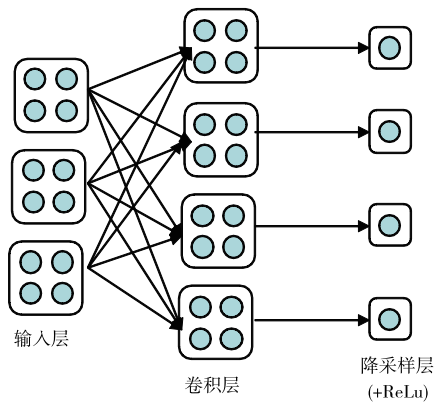

图 3 卷积神经网络的基本结构

Glorot[12]提出的深度稀疏校正神经网络从结构上仍然属于全连接神经网络,唯一变化是将激活函数替换成 ReLU. Barron[13]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. Hubel&Wiesel[14]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[15~19].

3 深度神经网络结构与超参数设计

3.1 深度卷积神经网络的设计

针对静态面部表情识别的任务,即从一个图像中识别出其中人物面部的表情,首先构建一个卷积神经网络. 如图 4,针对输入大小为 32x32 的灰度图(彩色维度为 1),构建了 3 个卷积 &MaxPooling 层,1 个全连接层,1 个 Softmax 层. 如图 4 所示. 根据各层神经元个数的不同,又分为:

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

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

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

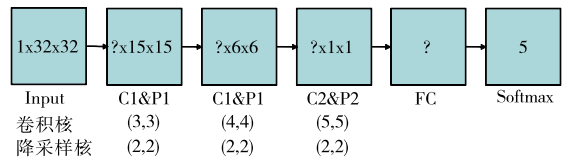

图 4 深度卷积神经网络的结构.?表示多种方案

3.2 深度稀疏校正神经网络的设计

为了对比验证,构建了一个深度稀疏矫正神经网络. 如图 5,针对输入大小为 32x32 的灰度图(彩色维度为 1),构建 3 个全连接层,1 个 Softmax 层.

根据各层神经元个数的不同,又分为:

DNN-1000:[1000,1000,1000]

DNN-2000:[2000,2000,2000]

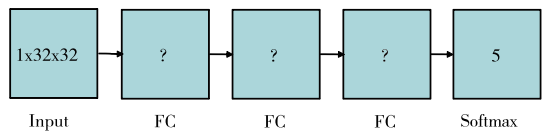

图 5 深度稀疏校正网络的结构.?表示多种方案

3.3 K 兴趣区域方法

Sun Y 等人[20]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. 以此为启发,根据人脸的面部结构,设置了 9 个不同的兴趣区域 ROI(Region of Interesting),如图 6,这 9 个区域也是直观上人判断情感的关注区域,通过设置这 9 个区域,主动引导神经网络关注与表情相关的面部区域.

兴趣区域分割方案重点关注眼、鼻、嘴在不同表情

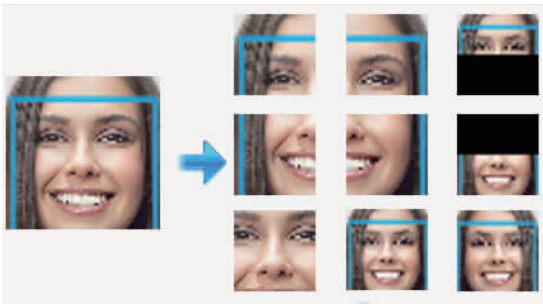

图 6　九个ROI区域

中区别,为了便于处理并保证方法的鲁棒性,没有预先检测面部关键点来分割,而直接设置固定区域分割. ROI 方法同时让训练数据扩大至 9 倍. ROI 数据扩大方案是针对训练阶段的图像,而在测试阶段可以直接对测试图像判别. 然而,这会浪费模型中记忆的关于 ROI 区域的分布式表达特征. 基于 K-ROI 方法对测试图像同样划分为 9 个区域,将 9 个区域视为原始图像的替代(即 K 取 9),对 9 个 ROI 区域的判别结果投票,取票数最多的结果作为最终结果.

3.4　样本旋转增强

Lopes 等[2]扩大数据集的方法是将原始图像旋转一定角度,生成大量角度变化的训练样本. Jaderberg[6]等人认为,对于数据预先做人工空间变换处理,将这一关注点嵌入至神经网络训练时,可以引导神经网络获得对应的不变性. 卷积神经网络的结构在对数据训练时,具有一层自适应变换能力. 然而,针对面部表情的识别,旋转采样的方案并不一定适合,甚至有可能会有负面效应. 这点会在实验来进一步验证.

4　实验与对比

为验证模型的适应能力,要引入 Wild 数据解决 CK + 数据集过于规范的问题. 通过搜索引擎收集了 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收集了 1200 张学生信息照片,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. 训练集由 CK + 的高兴、悲伤、惊讶、愤怒各 700 张混合互联网的 Wild 图片各 200 张、以及“中性”的 900 张构成. 总共 5 类,每类 900 张图片. 测试集由互联网收集图片各 300 张混合“中性”的 300 张构成. 共计 5 类,每类 300 张图片. 使用三个深度神经网络模型做对比评估,分别为 3.1、3.2 设计两个深度神经网络 CNN 和 DNN,及将 3.2 中的深度从 3 降为 1 的浅层神经网络 1NN,实验包括:ROI 辅助,K-ROI 辅助,旋转样本增强.

4.1　ROI 辅助实验

ROI 辅助是实验评估重点,它反应着模型内部分布式表达的训练情况. 实验使用的是 5 类共 4500 张面部训练数据,5 类共 1500 张测试数据. 训练 4500 张数据经过 ROI 处理后,最终为 4500 * 9 = 40500 张,测试数据不做变化. 实验结果如表 1,基准为无 ROI,加“ * ”标记表示有 ROI. 其中“总体”是在所有样本上计算的错误率,系统的总体分类性能可以用该值评价. “推断”列是针对实验结果,做出的推测.

表 1　ROI 辅助的测试集错误率

	中性	高兴	悲伤	惊讶	愤怒	总体	推断
CNN-64	4.7%	37.3%	51.7%	35.0%	39.7%	33.7%	baseline
CNN-64 *	7.0%	36.3%	60.0%	23.7%	30.0%	31.4%	增强
CNN-96 *	4.3%	37.0%	59.3%	19.0%	29.7%	29.87%	\
CNN-128	3.0%	32.0%	50.7%	29.0%	37.7%	30.5%	\
CNN-128 *	6.0%	30.3%	55.0%	17.7%	22.7%	26.3%	增强
DNN-1000	3.0%	38.3%	65.7%	38.0%	37.0%	36.4%	baseline
DNN-1000 *	3.0%	41.3%	52.7%	28.7%	30.7%	31.3%	增强
DNN-2000 *	2.3%	46.0%	52.7%	23.3%	32.0%	31.3%	\
1NN-1000	5.3%	36.0%	57.7%	43.7%	37.7%	36.1%	baseline
1NN-1000 *	5.0%	38.7%	59.3%	32.3%	42.3%	35.5%	无效
1NN-2000	4.7%	39.0%	54.0%	39.3%	39.3%	35.3%	\
1NN-2000 *	7.3%	39.7%	55.0%	35.0%	38.0%	35.0%	无效

从整体实验结果来看,ROI 的引入对两个模型的各规模都有 4 ~ 5% 的精度提升,符合预期. 卷积神经网络随着规模的提升,效果也得到提升,达到最好的整体错误率 25.8%. 通过对各个表情分析,能够获得出一些结

论. 首先,是中性测试集相对于其它测试集,测试成绩过于优异. 这是有意如此设置:测试集里,只有中性集,没有使用 Wild 数据,中性集是与训练集较为相似的正规数据,这个结果验证了 Lopes 等人[2]在 CK + 数据及上的高准确率测试结果并不一定能说明模型拥有了良好的泛化能力. 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、愤怒的测试结果则接近. 在对浅层神经网络 1NN 测试当中,发现训练集错误率较于深度神经网络而言,并没有太大变化,即无论是 1NN-1000 还是 1NN-2000 均不存在因模型容量不足,而导致的欠拟合,由于 Dropout 的引入,交叉验证中也没有发现明显的过拟合. 另外,在深度神经网络中,具有广泛提升的 ROI 方法,在浅层神经网络中作用并不明显. 此结果符合“深度结构能够逐层抽象”的推测,基于深度的逐层抽象组合,在组合视觉特征上,具有良好的分布式表示的能力,而不是像传统分类器(1NN、SVM)一样,对于数据只是简单训练局部空间距离,最终导致泛化性较差.

4.2 K-ROI 辅助实验

K-ROI 辅助评估将考察 K-ROI 方案中的投票对结果的影响,即在 4.1 的基础上,将测试数据采用 ROI 方法划分,单个测试图像的决策通过对该图像 9 个 ROI 的判别投票来确定. 按照推测,该方案对模型内部的分布式表示有较高的要求. 实验结果如表 2,基准为 4.1 中的有 ROI,“ * ”标记表示 K-ROI. 对各个表情的结果进行分析,在深度卷积神经网络中,除了悲伤集外,其它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,悲伤集变差且幅度最大,其它测试集几乎无变化. 此结果表明 K-ROI 的投票机制对模型的泛化能力有很高的要求,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,可能是得益于内部的自适应关注点机制. 从整体实验结果来看,K-ROI 方案中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,反而使整体结果略微变差. 与之相反的是,浅层神经神经网络在 K-ROI 的投票之后,均得到了更好的结果. 也就是几乎是完全依赖于局部分布表示的模型,却在分布表示要求高的任务中取得了较好的结果. 推测可能原因如下:因为对人脸表情 ROI 区域的划分非常正确,让其锁定了几个 ROI 区域,并且在多轮迭代中进行了过度地拟合、记忆了大量的局部距离信息. 如果该推测是正确的,那么这种过拟合现象会在引入旋转数据之后得到验证.

表 2 K-ROI 辅助的测试集错误率

	中性	高兴	悲伤	惊讶	愤怒	总体	推断
CNN-64	6.3%	39.3%	61.0%	26.7%	31.7%	33.0%	baseline
CNN-64 *	1.0%	35.6%	62.6%	22.7%	28.7%	30.1%	增强
CNN-96	4.3%	37.0%	59.3%	19.0%	29.7%	29.9%	baseline
CNN-96 *	1.0%	27.0%	60.3%	14.0%	28.7%	26.2%	增强
CNN-128	6.0%	30.3%	55.0%	17.7%	22.7%	26.3%	baseline
CNN-128 *	1.3%	24.3%	53.3%	13.7%	23.0%	23.1%	增强
DNN-1000	3.0%	41.3%	52.7%	28.7%	30.7%	31.3%	baseline
DNN-1000 *	0.0%	36.6%	59.0%	29.3%	30.0%	30.9%	无效
DNN-2000	2.3%	46.0%	52.7%	23.3%	32.0%	31.3%	baseline
DNN-2000 *	0.3%	42.0%	64.3%	25.0%	32.0%	32.7%	无效
1NN-1000	5.0%	38.7%	59.3%	32.3%	42.3%	35.5%	baseline
1NN-1000 *	1.0%	40.0%	58.0%	32.0%	33.7%	32.9%	增强
1NN-2000	7.3%	39.7%	55.0%	35.0%	38.0%	35.0%	baseline
1NN-2000 *	2.3%	40.7%	58.0%	33.3%	34.3%	33.7%	增强

4.3 旋转生成样本实验

在前面实验,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,为了验证该推测,制作了两份新的训练数据:

(I)针对 CK + 与学生照片两类正规数据,以图像中心为原点,进行旋转采样[2]. 对源训练集 5 类(每类 700)执行高斯随机数 11 次,加上 4500 训练图像,共有 $5 * 700 * 11 + 4500 = 43000$,构成新训练集,测试集不变化.

(II)将 I 中的 43000 个图像,与 4.1 部分中的 40500 个混合,共计 83500 个训练数据,构成新训练集,测试集不变化. 以 4.1 的无 ROI 测试结果作为对比基

准,实验结果如表3,加“ * ”标记表示使用I数据集,加“ + ”标记表示使用II数据集和ROI,加“^”标记表示使用II数据集和K-ROI方案.

表3　旋转样本增强的测试集错误率

	中性	高兴	悲伤	惊讶	愤怒	总体	推断
CNN-128	3.0%	32.0%	50.7%	29.0%	37.7%	30.5%	baseline
CNN-128 *	4.7%	42.0%	58.7%	32.0%	39.3%	35.3%	过拟合
CNN-128 +	3.7%	31.0%	54.7%	19.3%	25.3%	26.8%	修正
CNN-128^	0.3%	25.0%	59.7%	13.3%	24.0%	24.5%	增强
DNN-1000	3.0%	38.3%	65.7%	38.0%	37.0%	36.4%	baseline
DNN-1000 *	1.7%	41.0%	62.0%	37.3%	43.3%	37.1%	过拟合
DNN-1000 +	2.3%	41.3%	57.7%	29.7%	34.0%	33.0%	修正
DNN-1000^	0.7%	42.3%	69.0%	32.0%	32.3%	35.3%	无效
DNN-2000 *	2.7%	42.3%	65.3%	37.3%	44.7%	38.5%	过拟合
DNN-2000 +	2.0%	41.7%	60.0%	29.7%	32.7%	33.2%	修正
DNN-2000^	0.0%	37.0%	71.7%	35.3%	33.3%	35.5%	无效
1NN-1000	5.3%	36.0%	57.7%	43.7%	37.7%	36.1%	baseline
1NN-1000 *	4.3%	44.3%	59.3%	43.3%	42.7%	38.8%	过拟合
1NN-1000 +	4.7%	42.0%	58.7%	32.0%	39.3%	35.3%	过拟合
1NN-1000^	1.7%	41.3%	63.0%	28.7%	37.7%	34.5%	无效
1NN-2000	4.7%	39.0%	54.0%	39.3%	39.3%	35.3%	baseline
1NN-2000 *	4.7%	41.7%	59.7%	42.3%	47.0%	39.1%	过拟合
1NN-2000 +	4.3%	41.3%	60.3%	32.3%	40.0%	35.7%	过拟合
1NN-2000^	1.3%	44.3%	73.0%	27.0%	44.0%	37.9%	无效

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I数据集,CNN – 128、DNN – 1000用43000张原始与旋转生成的混合大数据,得出了比4500的小数据还差的结果,说明38500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等人[2]的结果截然相反,可能是基于CK + 的测试集与训练集的高度相似性掩盖了模型的过拟合问题. 其次,对于II数据集,ROI的引入几乎抵消了旋转样本的影响,但是此时K-ROI对于深度模型的效果普遍不佳,在DNN-1000中尤为明显. 结合前面的推测,可能引入旋转生成样本对分布式表示产生影响.

再次结合1NN-1000和1NN-2000,发现之前表现很好的K-ROI方案性能出现了严重降低,如果除去中性异常集,那么结果更差的. 即便不进行K-ROI投票,ROI与旋转对训练样本的共同强化,甚至比完全没有这些处理还要差. 根据之前的推测“因为对人脸表情ROI区域的划分非常正确,让其锁定了几个ROI区域,并且在多轮迭代中进行了强化拟合、记忆了大量的局部距离信息”,此实验中,引入旋转生成样本可能也对局部表示产生了影响. 通过DNN-2000的实验结果,结果发现DNN-2000基本和DNN-1000结果近似,排除了数据扩大2倍,产生欠拟合的可能性.

基于以上两个数据集的测试,可以得到引入旋转生成样本来扩大数据集并不一定适合面部表情识别任务. 它并不能让深度卷积神经网络获得很好的旋转不变性,反而因为旋转输入空间的引入,对缩小、平移不变性的效果产生干扰,导致的过拟合. 这种过拟合不是由于参数空间过大引起的. 当测试数据与训练数据有较大偏差和变化时,便会暴露.

5　基于实验结果的深入分析

5.1　深度假设

对比4.2节和4.3节稀疏校正神经网络在深度上的差异,基于实验结果重新分析了深度、局部表示、分布式表示之间的关系. 如图7所示,神经网络中间层,通过子函数复用、组合等原理,当受到反向传播误差驱动时,便可挖掘数据中的分布式特征. 分布式特征让学习焦点从单像素迁移到了分布式像素群,该像素群可以呈任意不规则形状. 若从图模型观点来看,设像素为

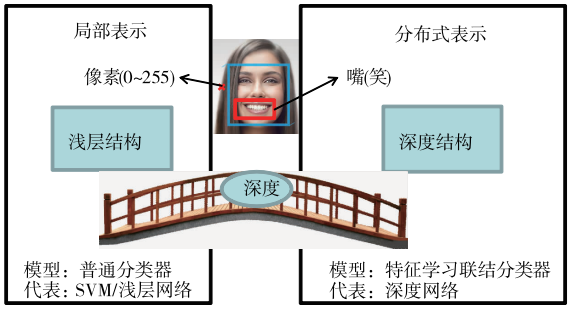

图 7 神经网络中存在的“特征桥”

基本单位,那么分布式表达特征,便可以从多个连通子图中挖掘信息. 如果具有相近表达行为的结点是同质的. 同质的结点之间可看作是强连通的,因为它们可互相访问交互与增强.

当图中存在强连通分量时,意味着结点存在冗余(等价于特征的线性相关),可以通过归并的方式合并结点. Tarjan[21]给出了通过深度搜索、标记结点与分量的做法,求解图中的强连通分量. 神经网络的反向传播、神经元激活机制,与 Tarjan 在离散图中求解强连通分量算法在一定程度上是等效的,如图 8. 在神经网络的反向传播过程中,梯度流会流过各层网络,通过激活函数这个阀门系统,或抑制,或激励. 该过程可等效于 Tarjan 算法在每次深度搜索时,对结点的分类标记. Tarjan 算法每次选取一个非标记结点而进行多轮深度搜索,且由于每轮深度搜索时,整张非连通图的标记情况都在变化,在神经网络的正向传播过程中,同样可以依赖前一层对后一层产生影响为参照来等效.

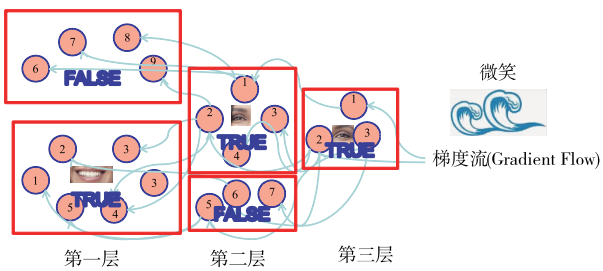

图 8 与Targin算法类似的基于连续型值求解连通分量方法

综合考虑正向传播与反向传播的等效原理,神经网络的深度可类比于 Tarjan 算法在图中进行深度搜索的次数. 当 Tarjan 算法完成对所有结点的搜索时,会得到唯一、且确定的连通分量组. 反之,只能得到局部、不确定的连通分量组. 迁移至神经网络中,当神经网络深度超过确定临界值时,将有可能在给定充足样本下,获得训练出全部特征组的基本条件. 反之,根据深度的减少,会在获得特征组数量的上界上,得到等量的削减. 更进一步,可推测出一种深度(depth)、局部表示(LR)、分布表示(DR)之间的关系:

$$\begin{cases} \text{capability}(\text{DR}) = \alpha \cdot (\text{depth} - 1) \\ \text{capability}(\text{DR}) + \text{capability}(\text{LR}) = 1 \\ \alpha \in (0,1) \end{cases} \tag{1}$$

在式(1)中,DR 和 LR 的容量总和是固定的,DR 容量与深度值成正比关系,为正比系数,范围为(0,1). 单隐层神经网络的 DR 容量为 0,LR 容量为 1. 而多隐层神经网络,随着深度的增加,DR 容量会越来越大,LR 容量会越来越小. DR 和 LR 的容量代表了它们在神经网络中的重要性. LR 的容量大于 DR,则对数据更容易产生局部距离学习. DR 的容量大于 LR,则对数据更容易产生分布式学习.

5.2 深度假设的实验假设

在面部表情识别中,提出一个比较深度神经网络局部表示与分布式表示的方法:

(1)左眼对左眼的强化,可用于评估局部表示.

(2)左眼对上半脸的强化,可用于评估分布表示.

(1)和(2)看起来类似一对互斥量. 从局部空间距离来看,微笑的左眼对愤怒的左眼响应系数可能是大于微笑的左眼对微笑的上半脸的. 从分布式空间距离来看,微笑的左眼与微笑的上半脸距离紧密,但微笑的左眼同愤怒的左眼,可能相距甚远. 这样的互斥性,最直接体现在 4.1 节与 4.2 节的实验结果中,DNN 和 1NN 之间,在 ROI 和 DNN 两种结果上的博弈与权衡:在 4.1 节中,ROI 效果在 DNN 上超过 1NN. 在 4.2 节中,K-ROI 效果在 1NN 上超过 DNN. 这两种互斥的结果,本质上可能缘于神经网络对 ROI 区域产生了不同的理解. 基于此,根据情感特征在一副人脸图像的分布可能性,切分了相应的 ROI 区域,而神经网络则可能存在两种途径去理解注入的这两种先验知识:

(1)狭义:这是对于局部空间距离的直接判断.

(2)广义:这是对于全局空间距离的间接判断.

ROI 效果在 1NN 上不佳,是因为在 1NN 中,capability(LR) > capability(DR),对于 ROI,产生了狭义角度的理解,在 K-ROI 上,对于这些局部分块的判断效果很好. ROI 效果在 DNN 上表现较好,是因为在 DNN 中,capability(DR) > capability(LR),对于 ROI,产生了广义角度的理解,在 K-ROI 上,反而对于局部分块的判断产生模糊. 这种基于相同结构、不同深度上的互斥结果,从一定程度上支持了式(1)的假设.

5.3 不变性假设

4.1 节中,DNN-1000 与 1NN-1000 这两个模型在没有 ROI 辅助情况下,结果几乎是相近的. 这似乎与 5.2 节中对深度的优势性分析产生矛盾. 为解释该现象,为 DR 定义一个新属性:ability(DR),它代表深度结构对于分布式特征的构造能力. 需要区别的是,ability(DR)与 capability(DR)是两个完全不同的概念. 如 5.1 节中

所述:

$$\text{capability(DR)} \propto \text{depth}$$

假设深度最终只会影响获得特征组数量的上限,而不是去引导神经网络如何获得特征组. 深度学习中的关注点机制可以引导获得特征组,显然,对于仅仅提升深度而言,并不会认为类似于一种关注点. 与其相反的是,深度的提升更像是一种容量的提升,一种对于逐层抽象、获取复杂关注点的必要条件的满足. ROI 在满足这种条件后才产生巨大的差异的情况,可以从一定程度上证明这一假设的合理性.

在广义上,ability(DR)来自于模型对于数据分布中重要特征组的关注. 对于图像数据,应当关注那些最能反映表达内容的图像块,对于文本数据,则应当关注那些最能体现语义的词与短语. 特别的,针对图像数据,针对图像中不变性的关注,最有可能获得效果极佳的、符合图像图义的特征组,直接提升 ability(DR),并且应用于 capability(DR). 鉴于图像的空间变换会给数据带来破坏,可以采用空间变换获取不变性的策略,定义 ability(T),且有:

$$\text{alility(DR)} \propto \text{ability(T)}$$

即用 ability(T)替代 ability(DR). 其中 T 表示经过空间变换后的数据分布表示.

5.4　不变性假设的实验解释

为了比较不同结构,首先需要排除 CNN 在实验中过度学习的可能性. DNN-1000 的参数量是 CNN-128 的 10 倍,DNN-2000 的参数量是 CNN-128 的 40 倍,理论上 CNN-128 更有可能出现欠拟合,但实际上它却达到了最好的效果. 4.2 节实验中,具有同样深度、不同结构在 CNN 与 DNN 呈现出较大差异. CNN 的各规模,无论在 ROI、还是 K-ROI 条件下,均表现出了良好的适应性,这可以从一定程度上证明深度不是决定分布式表示最终效果的唯一因素. 对比 5.2 节,还要考虑另外一个问题:当 capability(LR)被弱化之后,CNN 是如何提升左眼对左眼的学习效果? 为了回答该问题,再次定义 effect(DR),effect(LR),分别表示 DR、LR 的最终期望效果,有:

$$\begin{cases} \text{effect(DR)} = \\ \text{ability(T)} * \text{capability(DR)} + \text{capability(LR)} \\ \text{effect(LR)} = \\ \text{ability(T)} * \text{capability(LR)} \end{cases} \tag{2}$$

考虑到 DR 和 LR 两者互斥独占,将 ability 作为系数与之相乘. 另外,由于深度作为一种附加的物理结构,可以推测即便 DR 在神经网络中独占,仍然会受到 LR 的部分影响,所以在 effect(DR)中设置了一个偏置项,而 effect(LR)则无需设置该项. 考虑 effect(DR),在深度相同的 CNN 和 DNN 当中,决定效果的是 ability(T)的值. 数据空间变换可以提升它,自适应空间变换也可以提升它,且按照 Jaderberg 等人的观点,自适应空间变换效果要好于预处理变换,尤其是针对旋转变换时,预处理旋转很大程度上会造成 ability(T)的值变差而过低,呈现 4.3 的结果. 利用式(1)化简 effect(DR),

$$\text{effect(DR)} = \text{ability(T)} + [1 - \text{ability(T)}] * \text{capability(LR)} \tag{3}$$

利用式(3)求解不等式 effect(DR) > effect(LR),有:

$$\begin{cases} \text{capability(LR)} < \dfrac{\text{ability(T)} - 1}{2 * \text{ability(T)} - 1} \\ \text{capability(LR)} = 1 - \alpha \cdot (\text{depth} - 1) \end{cases} \tag{4}$$

式(3)和(4)这两个式子能解释 5.2 节中,DNN 与 1NN 在 ROI 效果上差异. 据此推测,当深度过低时,effect(DR)有很大可能弱于 effect(LR). 当注入的 ability(T)(ability(T) > 1)越强,需要满足不等式的要求就越苛刻:对 capability(LR)的值要求越小、对深度的要求越大,符合 5.3 节中对 ability(DR)、capability(DR)之间关系推测.

5.5　综合假设

综合考虑 5.1、5.3 节中假设,构造了图 9 中的坐标系,为关注点与不变性单独设置了空间轴,它表示模型对于数据理解的强度. 深度轴与强度轴的交叉,这一平面代表模型是卷积神经网络 CNN,以及 Jaderberg[6] 等人提出的空间变换网络 ST-NN. 它们本身对于数据分布中的关注点非常敏感,且深度结构保证了多层抽象、组合的物理条件. 深度轴与广度轴的交叉,代表模型是经典的 DNN. DNN 的全连接性,不适合也很难挖掘图像中的局部特征. 因为缺乏强度上的支持,它在的图像的分布式表示上,效果是非常弱的. 强度轴与广度轴的交叉,代表模型是经典的 SIFT + SVM. SIFT 的尺度不变特征变换,从很大程度上的缓解了 SVM 这样传统分类器的拟合压力. SVM 可看作是深度为 1 的特殊神经网络. 由于深度固定,支持向量数量将在训练中持续增长. 在大多数情况下,SVM 的广度(即拟合参数量)要比卷积

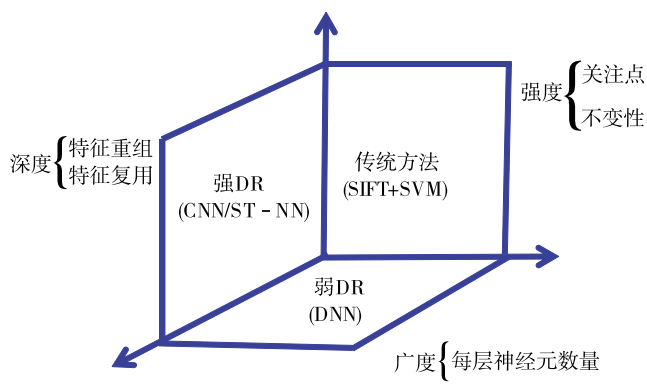

图 9　深度、广度、强度及其组合模型

神经网络大的多,所以,这类模型可看作是在广度和强度轴上的无限增长.

5.6 深度神经网络展望

基于5.5节中假设,当前深度神经网络的潜力可能体现在如下两点:(1) 由于神经网络特殊的联结结构,使得人工先验知识和方法可以大量嵌入. 借助于逐层贪心预训练、反向传播等机制,可以很自然的融合运用,短时间内就能让深度神经网络的效果倍增,无须重复计算人工设计的特征. (2) 由于深度的增长,在可控的条件下,模型的抽象、组合、泛化能力在多层叠加后,可能呈现指数级的提升. 当把(1)、(2)结合起来考虑,深度神经网络便可以变得十分强大,理论上,当样本充足时,可以超过任意的人工、传统方法,这可能会成为未来神经网络的重点发展方向.

6 结论

本文所提出的模型所学习到的特征,要优于传统的人工设计的特征[22]. 这是因为在设计深度神经网络的时候,一方面引入了面部表情的结构先验特征,另一方面调整深度神经网络,去掉其中不适合面部表情识别的特征处理方式,从而获得更优的深度学习模型. 基于此提出的 K-ROI 方法,对于面向面部表情识别的深度神经网络,有普遍优异提升效果. 更深层次的,还可以间接利用该方法分析模型中分布表示的情况,避免了完全的黑盒,实现了深度神经网络的一定程度上的"灰度化". 此外,深度与不变性将有可能成为增强图像神经网络最快速、最有力的工具. 通过对比实验,验证了两者在结构与功能上可能是互补的.

参考文献

[1] Alex K, Ilya S, Geoffrey H. Imagenet classification with deep convolutional neural networks[A]. Advances in Neural Information Processing Systems[C]. Lake Tahoe: Bartlett P, et al, 2012. 1097 – 1105.

[2] Andre T L, Edilson D A, Thiago O S. A Facial expression recognition system using convolutional networks[A]. 28th SIBGRAPI Conference on Graphics, Patterns and Images [C]. Bahia, Brazil: IEEE, 2015. 273 – 280.

[3] Lucey P, Cohn J, Kanade T, Saragih J, Ambadar Z, and Matthews I. The extended cohn-kanade dataset (CK +): A complete dataset for action unit and emotion-specified expression[A]. Conference on Computer Vision and Pattern Recognition Workshops [C]. San Francisco, US: IEEE, 2010. 94 – 101.

[4] Christopher M B. Pattern Recognition and Machine Learning[M]. New York: Springer, 2006.

[5] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1 – 127.

[6] Max J, Karen S, Andrew Z, Koray K. Spatial transformer networks[A]. Conference on Neural Information Processing Systems[C]. Montréal, Quebec, Canada: Curran Associates Inc, 2015. 2017 – 2025.

[7] LeCun Y, Boser B, et al.. Handwritten digit recognition with a back-propagation network[A]. Conference on Neural Information Processing Systems [C]. Morgan Kaufman, Denver: Morgan Kaufmann, 1990. 396 – 404.

[8] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. Biological Cybernetics, 1980, 36(4): 193 – 202.

[9] Rumelhart D E, Hinton G E, , Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(9): 533 – 536.

[10] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278 – 2324.

[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[DB]. arXiv preprint arXiv: 1409. 4842, 2014.

[12] Xavier G, Antoine B, Yoshua B. Deep sparse rectifier neural networks[J]. JMLR W&CP, 2011, 15: 315 – 323.

[13] Andrew R B. Universal approximation bounds for super positions of a sigmoidal function[J]. IEEE Trans on Information Theory, 1993, 39(3): 930 – 945.

[14] Hubel D H, Wiesel T N, LeVay S. Visual field representation in layer IVC of monkey striate cortex[A]. Society for Neuroscience, 4th Annual Meeting[C]. St Louis, US, 1974. 264 – 265.

[15] Dayan P, Abott L. Theoretical Neuroscience[M]. MA: MIT Press, 2001.

[16] Attwell D, Laughlin S. An energy budget for signaling in the grey matter of the brain[J]. Journal of Cerebral Blood Flow and Metabolism, 2001, 21(10): 1133 – 1145.

[17] Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors [DB]. arXiv preprint arXiv: 1207. 0580, 2012.

[18] Charles D. On the origin of species[R]. UK: John Murray, 1859.

[19] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[A]. International Conference on Artificial Intelligence and Statistics[C]. Italy: Sardinia, 2010. 249 – 256.

[20] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10000 classes [A]. Conference on Computer Vision and Pattern Recognition [C]. Colum-

bus,USA:IEEE,2014. 1891 – 1898.

[21] Tarjan R E. Depth-first search and linear graph algorithms [J]. SIAM Journal on Computing, 1972, 1(2): 146 – 160.

[22] 潘泓,朱亚平,夏思宇,金立左. 基于上下文信息和核熵成分分析的目标分类算法[J]. 电子学报,2016,44(3): 580 – 586.

PAN Hong, ZHU Ya-ping, XIA Si-yu, JIN Li-zuo. Object classification using context cue and kernel entropy component analysis[J]. Acta Electronica Sinica, 2016, 44(3): 580 – 586. (in Chinese)

作者简介

孙　晓　男,1980 年出生,山东龙口人,博士,副教授,硕士生导师. 研究方向为人机交互系统、情感计算与机器学习.

E-mail:sunx@ hfut. edu. cn

潘　汀　男,1995 年出生,江苏连云港人,本科生. 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