计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(4):55−63 [doi: 10.15888/j.cnki.csa.009825] [CSTR: 32024.14.csa.009825]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于微表情特征的谎言识别①

陈灿鹏[1], 吴桂兴[2], 郭　燕[2], 李春杰[2]

[1](中国科学技术大学 软件学院, 合肥 230026)
[2](中国科学技术大学苏州高等研究院 软件学院, 苏州 215123)
通信作者: 吴桂兴, E-mail: gxwu@ustc.edu.cn

摘　要: 目前, 有多种谎言识别方法, 包括使用测谎仪测谎. 然而这些方法执行起来效果有限, 不仅需要与被测谎对象产生接触, 而且要求相关人员具备专业知识, 不便于实行, 且效果有限. 心理学研究表明, 微表情是人脸上的一种持续时间极其短暂的细微肌肉运动, 能反映人在做出此表情时的真实内心状态. 相关研究表明, 人脸上的微表情特征可以作为谎言识别的线索. 本文研究基于微表情特征的谎言识别, 首先构建一个说谎时的微表情数据集, 命名为 MED. 其次, 设计一个基于多层自注意力机制的微表情特征学习模型 MEDR, 根据学习到的说谎和未说谎时的微表情特征进行谎言识别. 最后, 本文还在新构建的数据集上, 对本文设计的模型与一些现有模型进行实验对比, 实验结果显示, 本模型在自制高质量数据集上取得 94.33% 的准确率, 表明本模型在谎言识别方面具备出色的性能.

关键词: 深度学习; 计算机视觉; 表情识别; 微表情特征; 谎言识别

引用格式: 陈灿鹏,吴桂兴,郭燕,李春杰.基于微表情特征的谎言识别.计算机系统应用,2025,34(4):55–63. http://www.c-s-a.org.cn/1003-3254/9825.html

Deception Recognition Based on Micro-expression Features

CHEN Can-Peng[1], WU Gui-Xing[2], GUO Yan[2], LI Chun-Jie[2]

[1](School of Software Engineering, University of Science and Technology of China, Hefei 230026, China)
[2](School of Software Engineering, Suzhou Institute of Advanced Study, University of Science and Technology of China, Suzhou 215123, China)

Abstract: Currently, there are various methods for identifying lies, including the use of lie detectors. However, these methods have limited effectiveness in execution, as they not only require contact with the subject being tested for lies but also require relevant personnel to possess professional knowledge, making them inconvenient and less effective. Psychological research shows that micro-expressions are subtle muscle movements on the face with an extremely short duration, which can reflect a person's true inner state when they occur. Related studies show that micro-expression features can serve as clues for deception recognition. This study focuses on deception recognition based on micro-expression features. Firstly, a dataset called MED, which contains micro-expression data when people are lying, is constructed. Secondly, a micro-expression feature learning model named MEDR based on a multi-layer self-attention mechanism is designed. It can recognize lies based on the learned micro-expression features in both lying and non-lying situations. Finally, experimental comparisons between the proposed model and some existing models are conducted on the newly constructed dataset. Experimental results show that the proposed model achieves an accuracy of 94.33% on the self-made high-quality dataset, indicating its excellent performance in deception recognition.

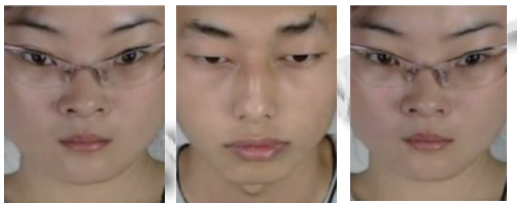
Key words: deep learning; computer vision; expression recognition; micro-expression feature; deception recognition

① 基金项目: 江苏省自然科学基金面上研究项目 (BK20141209)
收稿时间: 2024-09-23; 修改时间: 2024-10-08, 2024-11-12; 采用时间: 2024-11-18; csa 在线出版时间: 2025-03-04
CNKI 网络首发时间: 2025-03-06

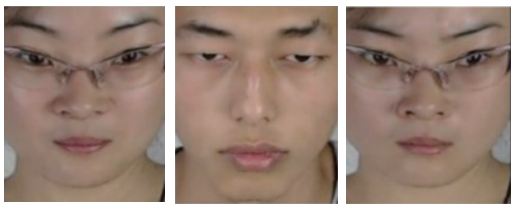
© 中国科学院软件研究所 http://www.c-s-a.org.cn

1 引言

1969 年, 心理学研究专家 Ekman 在研究谎言的过程中发现了微表情[1]. 微表情指在一段相当短暂的时间内 (一般在 0.04–0.2 s 之间) 出现在人脸上的细微的、难以立即察觉到的表情变化. 这种表情变化是人在短暂时间内情绪和内心状态的非自控且不易察觉的表达, 通常能够反映个体真实的情感和意图, 往往在人们试图隐藏或控制情感的时候出现, 因此可以作为谎言识别的有效行为线索[2]. 与微表情相对的是宏表情, 一般会在面部持续大约 0.2–4 s, 面部肌肉运动比较明显, 并且覆盖的脸部区域更大. 宏表情与微表情的对比如图 1 所示.



(a) 微表情



(b) 宏表情

图 1 微表情与宏表情对比 (图片来自 CAS(ME)^2 数据集)

Ekman 等人[3]于 1974 年提出了短暂表情识别测验 (brief affect recognition test, BART). 在后续的实验中他们发现被试者的微表情识别能力与谎言识别能力成正相关[4]. 近年来也有研究表明, 个体在说谎时往往会有一个认知负荷的变化过程, 此时的眨眼行为相当于反映认知负荷变化的指示器, 可以作为识别谎言的线索[5]. 但是, 如果仅依靠人工手段来识别微表情, 是一件极为耗时且准确率低的事情[6]. 因此, 使用计算机视觉的方法进行识别成为一个不错的选择.

目前, 在谎言识别领域已有一些数据集可供研究, 例如 Real-life trial data、Bag-of-Lies 等. 然而这些数据集的样本均来自外国人种, 并且存在画面质量不佳、样本数量极少等问题, 更为重要的是大部分数据集通过公开渠道获取十分困难. 这些因素都给基于微表情的谎言识别研究造成一定困扰.

基于以上研究观点, 为了填补该领域亚洲人脸构成的数据集的空白、提供可用的高质量数据集, 也为给谎言识别提供更多的可行方法, 本文首先构建了一个基于亚洲人脸说谎时的人脸微表情数据集, 接着提出了一种基于深度学习、借助微表情特征来识别谎言的方法.

本文的主要贡献如下.

(1) 提出了本文的研究目标之一, 即构建一个专用于识别谎言的亚洲人脸微表情数据集. 数据集命名为 MED, 其中包含 600 个视频片段, 及对各个数据样本进行标识的文件. 这为本文的研究, 更为日后相关领域研究者们提供了一个质量相对较高的开源数据集.

(2) 提出了一个基于微表情特征进行谎言识别的方法, 命名为 MEDR, 并且引入了相关实验进行验证和分析, 证明了本模型在谎言识别方面取得的卓越表现, 其效果超越了现有的基准模型.

2 相关工作

2.1 谎言识别

近年来, 非接触式的谎言识别成为一个重点研究领域. 当前大部分谎言识别模型使用了多模态进行综合分析. 例如, Pérez-Rosas 等人[7]研究出一种使用语言和非语言特征进行自动谎言识别的方法, 融合两种特征分析实现了 82% 的分类精度. 2018 年, Wu 等人[8]提出了一种视频中的谎言检测方法, 该方法综合了视觉、音频和文本、注释等不同模态来进行谎言识别, 实现了最高 0.922 的 AUC (area under the curve), AUC 是评估分类模型性能的指标, 表示 ROC 曲线下的面积, 值越接近 1 表明模型性能越好. 然而, 此类方法仍有不足之处, 例如, 使用的数据集质量不高、容易给模型引入过多的噪声从而导致模型识别效果不佳, 并且多模态方法需要音频和文字记录提供信息补充等. 在综合了已有研究方法并研究了微表情用于谎言识别的相关理论后, 本文研究一种基于微表情特征进行谎言识别的方法, 并提出了一个全新的高质量数据集.

2.2 对比学习

对比学习作为一种无监督学习方法, 近年来广泛应用于图像分类等计算机视觉任务中. 其核心思想是通过比较数据样本之间的相似性和差异性来学习数据的特征表示. 在计算机视觉中, 常见的方案是对统

© 中国科学院软件研究所 http://www.c-s-a.org.cn

一图像进行多次随机变换 (例如, 裁剪、反转、扭曲和旋转等) 来生成相似数据对[9]. 对比学习的核心是将两个输入数据分别映射到两个嵌入向量, 让这两个嵌入向量共享相同的模型参数, 通过计算这两个嵌入向量的距离, 可以得到他们相似度的量化结果. 对比学习的关键在于损失函数, 需要尽可能缩小相似输入之间的距离, 拉远不相似输入的距离. 目前, 一个常用的对比学习损失函数是 triplet loss[10], 通过构造一个锚样本、一个正样本和一个负样本的三元组来训练对比学习模型. 但 triplet loss 的训练过程比较困难, 容易产生过拟合现象和平凡解 (所有样本嵌入向量都趋向于 0) 问题. 因此, 有研究人员提出了 InfoNCE 方法[11], 该方法使用自信息最大化 (InfoMax), 借鉴了噪声对比估计并将对比学习转化为分类问题, 即给定一个锚样本和多个负样本, 判断每个负样本是否与锚样本相似, 使用相似度来衡量锚样本与正负样本的差距, 并且采用了更好的香农熵形式来计算相似度. 在本文的基于微表情特征进行谎言识别的工作中, 使用了对比学习进行模型训练.

3 谎言微表情数据集的构建

3.1 现有数据集对比

目前, 已经存在多个多模态数据集能够用于谎言识别研究. 本节对现有数据集分别进行介绍并进行综合对比.

3.1.1 Real-life trial data 数据集

该数据集发布于 2015 年, 是从国外的公开庭审录像收集而来的视频, 是采用语言及非语言的方式建立起来的一个多模态数据集. 该数据集内包含的视频片段分辨率大小并不统一, 包括 854×480、480×320 等多种格式. 该数据集易于获取, 可用于本次实验对比.

3.1.2 Bag-of-Lies 数据集

该数据集发布于 2019 年, 是一个在真实场景中收集的多模态数据集, 提供了音频、视频甚至脑电图等在内的多种形式数据, 数据集提供了 325 个数据, 真实 (163 个) 与谎言 (162 个) 数量大体相当. 目前通过公开渠道难以获取该数据集, 故本文实验未使用该数据集.

3.1.3 Box-of-Lies 数据集

该数据集发布于 2019 年, 也是一个多模态的数据集, 数据内容源自美国 NBC 电视台节目《吉米今夜秀》(The Tonight Show) 主持人与嘉宾在玩游戏时的欺骗性对话, 双方在猜对方描述物体对象时是否有说谎行为存在. 此数据集包含了面部以及语言行为, 并且对多模态交流进行了相应注释. 该数据集通过公开渠道尚难以获取, 故不参与此次试验对比.

3.1.4 Multimodal 数据集

该数据集发布于 2014 年, 同样是多模态数据集, 该数据集记录了在 3 种互不相同的欺骗场景中被测人员的生理、视觉、热反应的测量数据, 能够为研究人员从新的视角进行谎言识别研究提供新的思路. 然而, 该数据集为非公开数据集, 故也无法参与本次实验.

3.1.5 现有数据集对比总结

已知的各数据集对比如表 1 所示.

表 1 谎言数据集对比

属性	Real-life trial data	Bag-of-Lies	Box-of-Lies	Multimodal
受试者数	56	35	26	30
受试者组成	男35, 女21	男25, 女10	男6, 女20	男25, 女5
平均年龄	16–60	—	—	22–38
样本数	121	325	1 049	150
样本组成	真话60 谎言61	真话163 谎言162	真话187 谎言862	真话74 谎言76
RGB图像	有	有	有	有
头部运动	有	有	有	—
音频	有	有	有	有
收集方法	真实场景	真实场景	真实场景	假设场景
发布年份	2015	2019	2019	2014

3.2 数据集制作

3.2.1 视频数据制作

本文的视频数据主要来源于国内互联网上的知名在线视频平台优酷, 通过在该平台付费获得视频观看资质, 保证本数据集的数据来源遵守相关法律法规. 本数据集的视频片段来源于国内一系列演出质量极高的电视剧, 如电视剧《读心神探》, 剧中大量片段是警探根据嫌疑人的微表情判断其是否说谎, 对人物微表情、动作等的演绎十分到位, 非常逼真, 非常有利于将这些片段作为实验数据.

本文首先对影视剧进行整体录制, 然后将人物作真实陈述或者说谎的片段截取为数据样本, 并按照剧情中人物是否说谎做标签归类. 对于视频片段的分类标注, 本数据集样本由 3 个人分别进行标注, 并采取 Kappa 一致性检验确保标注的准确性.

3.2.2 Kappa 一致性检验

本文对于数据集由不同标注者标注的结果, 采用

© 中国科学院软件研究所 http://www.c-s-a.org.cn

Kappa 进行一致性校验. Kappa 一致性检验是指使用 Kappa 系数来检验两种分类结果是否一致, 该系数关注的是两个观察者对同一事物的观测结果或者同一观察者对同一事物的两次观测结果是否一致, 从而可以展示出由于机遇造成的一致性与实际观测的一致性之间的差别大小. Kappa 系数值越大, 说明两种分类结果一致性越高[12,13]. 标注 Kappa 的计算过程如式 (1) 所示:

$$k = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

其中, p_o 是观察到的总体一致性比例, 即所有评估者一致分类正确的比例, p_e 是每一类正确分类的样本数之和除以总样本数, 即总体分类精度. 对于本文的研究目标, 共有两个类别, 分别是真话和谎言, 分别用 0 和 1 表示类别, 则两类的真实样本个数分别表示为 a_0、a_1, 预测出来的每 1 类样本个数分别是 b_0、b_1, 样本总数为 n, 则 p_e 的计算如式 (2) 所示:

$$p_e = \frac{a_0 \times a_1 + b_0 \times b_1}{n \times n} \tag{2}$$

之后, 对 3 组标注结果两两进行 Kappa 一致性校验, 假设将参与标注工作的 3 人的标注结果记为 $Label(i)$, $Label(j)$, $Label(k)$. k_{ij} 表示 $Label(i)$ 和 $Label(j)$ 进行 Kappa 一致性检验得到的 Kappa 值. 当 $k_{ij} \geqslant 0.4$ 时, 将 $Label(i)$ 和 $Label(j)$ 与其 *Coefficient* 值 (置信度值) 进行下一步计算, 如式 (3) 所示:

$$Label' = \text{round}\left(\frac{\sum_{k=1}^{n} Label(k) \times Coefficient(k)}{n}\right) \tag{3}$$

其中, n 表示经过 Kappa 一致性检验筛选下的 *Label* 组个数, k 表示筛选出来的 *Label* 组标号, 对向量中每个元素进行四舍五入取整, 最后得到的 *Label'* 也是一个软标签, *Label'* 向量元素的值也在 0 和 1 之间.

Label 为单选标签, 将 3 组 *Label* 中对应于每一个视频样本的标注进行比较, 如果同一个样本对应的元素值相同, 直接选取该元素为最终元素值, 如果 3 组的结果不一致, 则选择存在两个元素相同的值作为最终元素值.

3.2.3 自制数据集概况

本文将自制数据集命名 MED (micro-expressions for deception detection). 数据集制作完成后, 本文使用早前在 Real-life trial data 数据集上训练得到的效果最佳的 GRU 模型在 MED 数据集上进行谎言识别实验, 实验结果显示在 Real-life trial data 数据集上训练得到的 GRU 模型在 MED 数据集上表现不佳, 分类准确率极低. 这也证实了构建本数据集的必要性——在欧美等外国人种面部数据上进行训练的模型可能不适用于 MED 这种由亚洲人面部数据构成的数据集, 本数据集能够填补该领域空白.

MED 数据集中每个视频的时长为数秒到数十秒, 整个数据集共包含 600 个视频, 合计总时长约为 170 min. 其中每个视频的内容为人物回答问题或者谈话时的片段. MED 的一些重要属性如表 2 所示.

表 2 MED 数据集属性

属性	数据
数据来源	我国表演质量极高的知名电视剧
受试者人数 (个)	男50, 女50
样本数 (个)	600
样本组成 (个)	真话300, 谎言300
分辨率	1920×1080
受试者人种	亚洲人 (中国人)
头部运动	有
构建年份	2023–2024

图 2 分别展示了 MED 数据集中说谎和讲真话的两个不同场景视频中截取得到的连续视频帧. 通过对比可以直观地看出, 试验对象在讲真话和说谎时的表情有明显不同. 说谎者通常会有紧张或者不自然的感觉, 眼神不自觉地躲闪等. 而讲真话者表情自然放松, 面部表情与本人的言辞一致, 特别是眼神稳定.

(a) 来源: Deception_96.mp4

(b) 来源: Truth_124.mp4

图 2 MED 数据集部分画面图像

© 中国科学院软件研究所 http://www.c-s-a.org.cn

4 基于微表情特征的谎言识别方法

4.1 算法设计

4.1.1 模型结构

本节提出的方法需要进行微表情特征的提取和学习. 相关文献研究表明, 自注意力机制能够很好地学习微表情特征[14,15]. 因此, 本节实验提出一个基于多层自注意力模型 MEDR 进行基于微表情的谎言识别.

图 3 展示了 MEDR 模型中的多层自注意力网络部分的结构, 每 1 层包括 Transformer 以及块聚合. Transformer 对每 1 个图像块单独实现自注意机制. 在低层级网络中, 将面部分为 16 个区域, 将细粒度特征进行聚合, 将小的图像块聚合成为 4 个更大的块, 从而减少数据维度. 接着将数据传递到高层自注意力网络, 对低层网络提取到的特征进行自注意, 捕捉粗粒度特征. 最后, 经过高层 Transformer 处理得到的特征图传递到最大池化层和全连接层进行谎言分类.

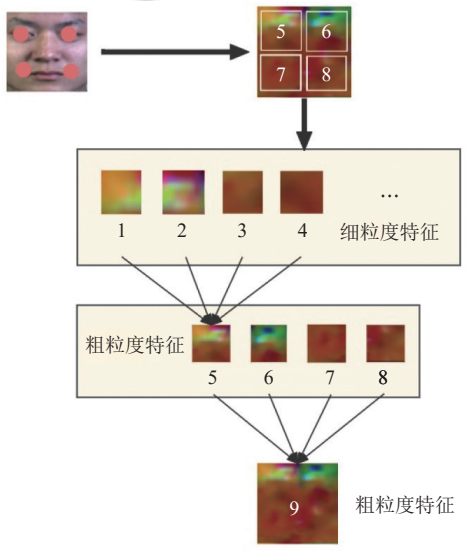


图 3 多层自注意力网络结构

综合相关研究, 本文提出的多层自注意力网络具有如下优势: 对于局部特征提取, 低层网络将面部分为 16 个区域, 每个区域专注于不同区域的面部特征, 使模型能够捕捉到更加细腻的表情变化, 同时将区域进行聚合, 可以减少数据的维度. 高层网络接收聚合后的特征进行进一步的自注意, 能够在更高的层次上理解面部表情的变化, 捕捉到更加复杂的模式. 相较于直接使用一个深层 Transformer 网络, 本文采用分层处理能够减少模型复杂度, 从而降低发生过拟合的风险, 特别是在训练样本有限的情况下. 同时, 两层 Transformer 结构的设计使模型能够根据不同数据和任务进行微调, 更容易适应多种多样的微表情特征. 这些优势印证了本模型能够基于微表情特征进行微表情识别.

如图 3 所示, 每 1 层自注意力层将面部分为多个区域, 第 1 层的网络将面部划分为 16 个区域, 以便于提取细粒度的面部微表情特征. 使用卷积核操作来实现这一划分, 每个卷积核大小为 12×12, 步长为 12.

如图 4 展示了模型第 1 层的局部特征学习网络结构, 其中a_i表示第i个局部区域向量, $i \in (1,2,\cdots,16)$. 通过 3 个可学习的矩阵W_q、W_k、$W_v \in \mathcal{R}^{3\times3}$投影它的潜在特征, 分别得到查询向量$Q_i$、键向量$K_i$以及值向量$V_i$, 计算过程如式 (4) 所示:

$$Q_i = a_i \times W_q,\ K_i = a_i \times W_k,\ V_i = a_i \times W_v \tag{4}$$

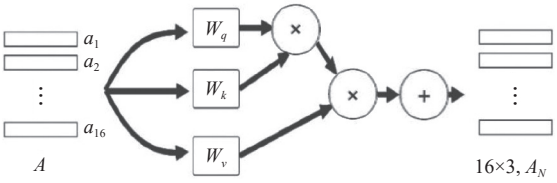


图 4 基于多层自注意力的微表情特征学习网络结构

将每个区域的查询向量和键向量的转置相乘, 从而得到区域间的相似度评分, 每个区域的值向量再通过对应的归一化值进行加权. 接着, 使用残差连接得到面部肌肉的局部节点表示, 计算过程如式 (5) 所示. 其中, 矩阵A、Q、K、$V \in \mathcal{R}^{16\times3}$分别表示$a_i$、$Q_i$、$K_i$、$V_i$按行合并, d_k表示矩阵K的行维度.

$$A_N = A + Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

局部区域特征学习网络如图 4 所示. 第 1 层 Transformer 处理上层聚合而来的 4 个区域, 在 4 个区域上进行自注意, 将 4 个区域的特征融合为一个整体特征.

由多层自注意力机制学习到的微表情特征代表了人物在说话时的面部肌肉运动主要信息, 以编码向量的形式传递到后续的全局分类模块, 该模块包含 2 个全连接层和 1 个 *Softmax* 层, 每个全连接层的作用是进一步提取特征并进行降维, 最终通过 *Softmax* 层将前面的输出再转化为分类概率以进行谎言分类.

4.1.2 多正例对比学习

目前多数对比学习方法都采用 Siamese 框架[16]来训练模型, 但该方法存在一些缺点: 1) 训练困难, 容易

© 中国科学院软件研究所 http://www.c-s-a.org.cn

出现过拟合情况. 2) 计算量大, 导致训练时间更长. 3) 泛化能力较弱, 对于未知的样本对, 预测效果较差.

基于以上问题, 本文改进设计另一种形式的对比学习——多正例对比学习. 与 Siamese 相比, 本文的多正例对比学习有以下优点: 1) 可以使用多个正样本和负样本, 从而更全面地捕捉样本之间的关系. 2) 可以学习到更具鲁棒性和泛化性的特征表示. 通过尽可能缩小相似输入之间的距离来优化模型. 3) 对抗性强. 训练时将相似的样本尽量聚集到一起, 与不相似的样本分开, 使得模型能够更好地对抗噪声影响. 4) 减少对数据集的依赖. 对比学习能在较小的数据集上进行训练, 并且不需要成对的样本, 使用场景更广泛.

对比学习最重要的在于为每个样本构造正例, 在基于微表情特征的谎言识别中, 每个视频的微表情帧之间构成正例, 同类谎言识别类型的不同视频微表情帧之间也构成正例. 在对比学习中, 输入为一个批次的样本视频连续帧嵌入向量 *Batch*, 并且已知批次内正负例关系, 设对于每个样本视频帧嵌入向量为s_i, $s_i \in Batch$, $R=\{\text{同类}, \text{不同类}\}$, $r_{i,j} \in R$, $(s_i, r_{i,j}, s_i)$表示这两帧中的微表情特征代表的谎言识别类型具有$r_{i,j}$关系, 同分类的不同视频的帧也能按照实际用此关系表示. 输出为对比学习训练损失$\mathcal{L}$.

对于多正例对比学习, 在同一批次内, 从说谎 (正例) 的视频中随机选择几个关键帧, 形成正例对, 记为g和g^+, 每个正例中的两个帧来源的样本都是说谎样本. 再从说谎 (正例) 视频和真话 (负例) 视频中随机选择关键帧形成负例对. 本文的多正例对比学习示意图如图 5 所示. 损失函数如式 (6) 所示:

$$\mathcal{L}=-\log\frac{1}{|S|}\frac{\sum_{k\in S}\mathrm{e}^{sim(g_i,g_k)/\tau}}{\sum_{j\in N}\mathrm{e}^{sim(g_i,g_j)/\tau}} \tag{6}$$

其中, S 为同一批次内样本 i 的正例集合, k 为 i 的一个正例. N代表同一批次内样本 i 的负例集合, j 为 i 的一个负例. $sim(g_i,g_j)$是g_i和g_j的余弦相似度, τ为温度参数. 对于每个样本 i, 设计的多正例对比学习通过同时拉近g_k (i 的正例) 和尽量推开多个g_j (i 的负例) 来进一步提高模型性能.

Siamese 的训练需要显式地为每个正例对指定一个负例, 相较而言, 本文设计的多正例对比学习对正例个数没有要求, 能够充分参考更多正例g^+的信息, 从而帮助模型更好找到每个同类微表情所具有的共同特征.

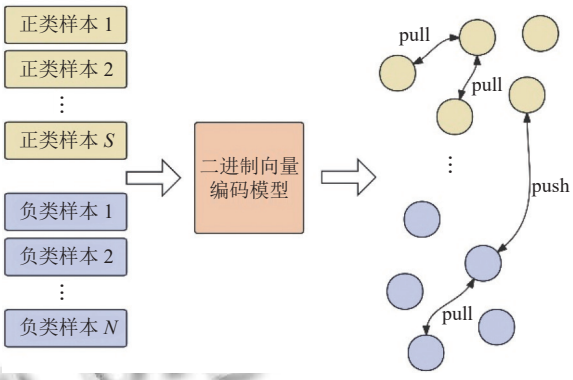


图 5 多正例对比学习示意图

4.2 实验与分析

4.2.1 实验设置

本节实验主要使用 PyTorch 深度学习框架, 优化器策略是 AdamW, 初始学习率为 0.000 2, 采取多卡 DDP 的训练策略, 每张卡设定 BatchSize=64, 表 3 说明了实验的具体环境.

表 3 实验环境

环境	版本号
CPU	Intel Xeon(R) Gold 6 133
GPU	NVIDIA GeForce GTX4090 × 4
内存	128 GB
操作系统	Ubuntu 20.04.6 LTS
CUDA	12.1
Python	3.10
PyTorch	2.1

4.2.2 实验过程

本节所使用的 MED 数据集中, 数据样本均为 mp4 视频格式. 为减少视频画面中背景噪声等干扰, 以及对数据集进行划分等操作, 需要先对数据进行预处理. 数据预处理的过程如下.

首先, 逐个读取所有视频, 对视频的每 1 帧进行处理, 消除多余的冗余帧信息, 针对每个片段中的微表情样本, 以峰值帧为中间帧, 向其前后两边展开, 选择合理的帧图片来统一微表情的帧长. 根据文献[17]的研究, 为降低计算成本, 在微表情样本中使用中间位置帧代替峰值帧进行微表情识别是合理的选择, 因此, 在此环节本文使用这一种方法来选择视频帧.

确定好帧图片后, 将每 1 帧的图像通过 OpenCV Dlib 库中的 68 点检测算法[18]来检测确定图中最贴合

© 中国科学院软件研究所 http://www.c-s-a.org.cn

人脸的矩形区域, 并进行裁剪, 使得图像中只包含人脸 (每一帧只有被测对象的脸会出现在图像中). 随后, 使用平面线性插值方法将每个裁剪后的面部图像转换为灰度和标准的 48×48 像素格式.

处理完毕后, 将数据存储为 1 个 numpy 数组. 整个数据集共有 600 个视频 (300 个谎言和 300 个真话), 先将这些数据按照 9:1 的比例划分成开发数据集和测试数据集, 然后将开发数据集也按照 9:1 的比例划分成训练数据集和验证数据集.

通过以上步骤, 可以将视频片段转换为由时间堆叠的人脸面部表情帧序列, 用相应的检测算法识别出微表情的峰值帧及其邻近帧并保存, 用于后续的微表情特征学习和谎言识别. 接下来, 将以 numpy 数组格式表示的面部图像读入内存, 然后使用面部微表情识别模型学习每个图像并转换为一个编码向量, 交由后续模块进行识别分类.

4.2.3 实验评价指标

本文研究的问题是判别人物是否说谎, 属于二分类问题. 因此, 本文采用最直观的混淆矩阵 (confusion matrix) 来衡量分类模型的准确度 (在本研究的谎言识别部分, 正例指的是说谎的情况, 负例指的是未说谎的情况), 二分类的各个参数解释如下.

True positive (*TP*): 真正例, 实际为正例 (positive), 模型预测也为正例.

False positive (*FP*): 假正例, 实际为负例 (negative), 模型预测为正例.

False negative (*FN*): 假负例, 实际为正例 (positive), 模型预测为负例.

True negative (*TN*): 真负例, 实际为负例 (negative), 模型预测也为负例.

由上述参数可计算得到如下的各个评估指标.

准确率 (*Accuracy*), 预测正确的结果占总样本数的比例:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

精确率 (*Precision*), 在所有被模型预测为正例的样本中, 实际为正例的比例:

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

召回率 (*Recall*): 在所有实际为正例的样本中, 预测为正例且实际是正例的样本所占比例:

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

*F*1-分数 (*F*1-*Score*), 综合了精确率和召回率的一个评估指标:

$$F1\text{-}Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{10}$$

4.2.4 基本对比模型

GRU: GRU 网络是 RNN 的一种变体形式, 用于处理和学习时间序列的数据. 本实验使用的是两层的单向门控循环单元, 当通过一维卷积层和最大池化层以后, 该网络可以用来探测时间序列数据中的模式. 根据实验结果, 该模型在 Real-life trial data 数据集上实现了 81.82% 的准确率[19], 足以用作基准模型进行实验.

CNN: 专门用于处理图像数据, 通过卷积层提取局部特征并进行特征学习. CNN 使用卷积运算和池化层减少数据维度, 并通过全连接层进行分类或回归. 它在图像识别、目标检测和其他视觉任务中表现出色, 是计算机视觉领域的核心技术.

3D-CNN: 直接使用原始预处理后的面部图像堆叠作为 Conv3D 模型的输入, 通过 8 层 Conv3D 将输入传递, 中间使用了最大池化层和 dropout, 然后通过 4 个全连接的密集层, 最后使用 *Sigmoid* 激活函数输出真/假的预测结果.

VGG16: 包含 16 层权重层 (13 层卷积层和 3 层全连接层). 它以小的 3×3 卷积核和 2×2 池化层构建深层网络, 能够提取图像的高级特征. VGG16 在 ImageNet 等大规模图像分类任务中表现优异, 并且在计算机视觉领域具有广泛的应用.

ResNet: 解决了深度网络训练中的梯度消失问题. 其核心思想是允许网络学习残差映射, 从而使得更深的网络结构能够更有效地训练. ResNet 显著提高了网络的性能和训练效率, 并在多个计算机视觉任务中取得了突破性成果. 常见的有 ResNet18 和 ResNet50.

4.2.5 实验结果及分析

由于 Bag-of-Lies、Box-of-Lies 以及 Multimodal 数据集目前无法获取, 本节分别在此前已有的 Real-life trial data 数据集和自制的 MED 数据集进行实验对比.

(1) 在自制 MED 数据集中的实验结果

表 4 展示了在 MED 数据集上将不同模型用于谎言识别的实验结果对比. 图 6 以混淆矩阵的形式展示了本文的 MEDR 模型在自制数据集上的识别结果.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

本文在训练模型时, 通过使用较小的初始学习率 (1×10^{-4}) 和预热 100 步后学习率线性衰减策略, 以及式 (6) 所示的对比学习损失函数来优化对模型的训练, 获得如图 7 所示的损失收敛曲线.

表 4 MED 数据集实验结果

模型	*Accuracy*	*Precision*	*Recall*	*F1-Score*
GRU	0.8867	0.8946	0.8767	0.8855
CNN	0.6800	0.6824	0.6733	0.6779
3D-CNN	0.6667	0.6678	0.6633	0.6656
VGG16	0.6917	0.7091	0.6500	0.6783
ResNet18	0.7417	0.7409	0.7433	0.7421
ResNet50	0.7617	0.7508	0.7833	0.7667
MEDR (Ours)	**0.9433**	**0.9926**	**0.8933**	**0.9404**

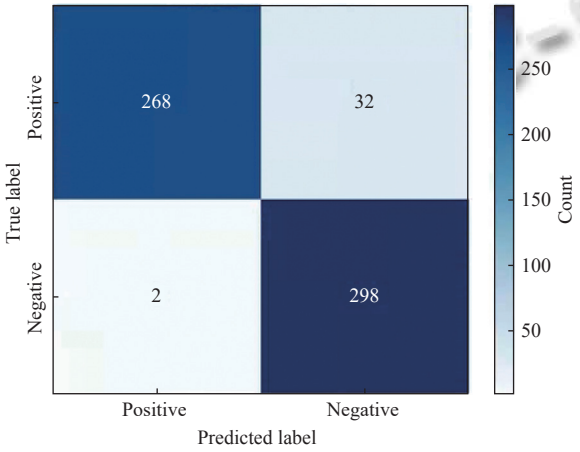


图 6 MEDR 模型在自制数据集实验结果

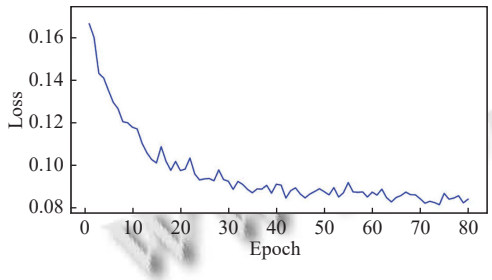


图 7 MEDR 模型在自制数据集损失收敛曲线

表 4 的实验结果表明, 本模型在谎言识别方面的性能已经能够达到当前的 SOTA 水平. 从实验得出的 *F1-Score* 指标可以看出, 模型的精确率和召回率均较高, 表明模型的误报和漏报情况较少. 同时, 从图 6 的混淆矩阵可以看出, 本模型对于近 90% 的谎言能够成功识别, 对于讲真话的情况, 仅有极少数被误判为谎言. 由此, 尽管本模型可能错过少部分谎言, 但将真话误判为谎言的可能性微乎其微, 这也是本模型的一个亮点.

(2) 在 Real-life trial data 数据集中的实验结果

表 5 展示了在 Real-life trial data 数据集上将不同模型用于谎言识别的实验结果对比. 图 8 通过混淆矩阵展示本文的 MEDR 模型在 Real-life trial data 数据集上的识别结果.

表 5 Real-life trial data 数据集实验结果

模型	*Accuracy*	*Precision*	*Recall*	*F1-Score*
GRU	0.8182	0.9149	0.7049	0.7963
CNN	0.6942	0.7222	0.6393	0.6783
3D-CNN	0.7025	0.7119	0.6885	0.7000
VGG16	0.6860	0.6949	0.6721	0.6833
ResNet18	0.6116	0.6129	0.6230	0.6179
ResNet50	0.6446	0.6557	0.6452	0.6504
MEDR (Ours)	**0.8595**	**0.9231**	**0.7869**	**0.8496**

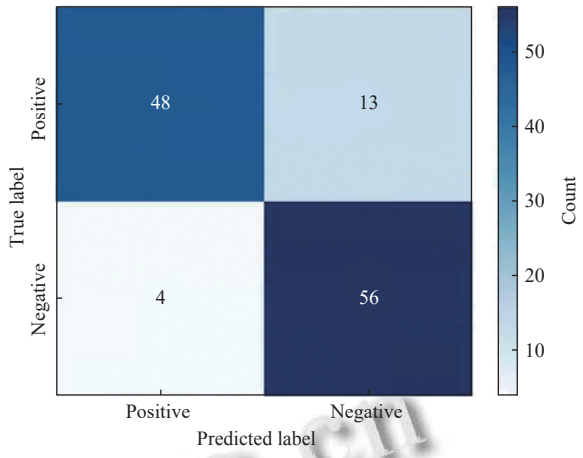


图 8 MEDR 模型在 Real-life trial data 数据集实验结果

从表 5 可以看出, MEDR 模型在 Real-life trial data 数据集上的各项指标仍优于 GRU 模型.

(3) 实验评价

由以上实验结果可以看出, 在自制的 MED 数据集上, 本文提出的 MEDR 模型实现了 94.33% 的准确率, 在国外现有的 Real-life trial data 数据集上, MEDR 模型仍实现了 85.95% 的准确率, 比文献[19]中的 GRU 模型表现更好, 表明本模型在基于微表情识别方面具有出色的性能.

5 结论与展望

为提供一种较为简便的基于计算机视觉的谎言识别方法, 本文首先构建了一个富含微表情细节的谎言视频片段数据集, 并通过 Kappa 一致性检验保证样本标注的可靠性. 接着提出了一种基于微表情特征的谎

© 中国科学院软件研究所 http://www.c-s-a.org.cn

言识别方法, 该方法采取多层自注意力机制学习人脸的微表情特征并基于此来进行谎言识别, 在模型训练中采用了多正例对比学习方法. 通过实验, 证实了本文设计的模型能够很好地捕捉到人脸的微表情信息, 并基于这些信息进行谎言识别, 取得了十分卓越的效果. 最后, 由于本文提出的方法是基于视频帧图片得到的整体面部微表情特征进行识别, 未来将考虑研究结合画面的时序变化以及人讲话时的身体动作进行谎言识别, 以更充分地利用视频片段的时间、空间序列特征. 同时, 未来还将进一步优化本文构建的数据集, 例如增加样本数量、进行更多维度的标记等.

在实际应用中, 本模型能够服务于刑侦讯问、心理健康等领域的谎言识别. 具体而言, 可以进行实时视频流分析, 采用高帧率摄像头、高效的面部检测与关键点识别技术等来捕捉面部的微表情变化, 同时也可以考虑使用模型剪枝和量化技术对本模型进行轻量化的优化, 这样使模型能够快速提取局部特征并进行实时判断.

参考文献

1 Ekman P, Friesen WV. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica, 1969, 1(1): 49–98. [doi: 10.1515/semi.1969.1.1.49]

2 吴奇, 李宇芳. 微表情知觉对人际信任的影响. 第二十四届全国心理学学术会议摘要集, 2022: 128–130. [doi: 10.26914/c.cnkihy.2022.070526]

3 Ekman P, Friesen WV. Nonverbal behavior and psychopathology. In: Friedman RJ, Katz MM, eds. The Psychology of Depression: Contemporary Theory and Research. Hoboken: John Wiley & Sons, 1974. 3–31.

4 Ekman P, O'Sullivan M. Who can catch a liar? American Psychologist, 1991, 46(9): 913–920.

5 石镇华. 谎言中的眨眼行为研究. 赤峰学院学报 (自然科学版), 2015, 31(22): 129–131. [doi: 10.13398/j.cnki.issn1673-260x.2015.22.053]

6 李文书, 张琛, 李宏汀, 等. 微表情识别方法综述. 人类工效学, 2018, 24(4): 75–80. [doi: 10.13837/j.issn.1006-8309.2018.04.0015]

7 Pérez-Rosas V, Abouelenien M, Mihalcea R, *et al*. Verbal and nonverbal clues for real-life deception detection. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 2336–2346.

8 Wu Z, Singh B, Davis L, *et al*. Deception detection in videos. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018.

9 Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 1597–1607.

10 Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 815–823.

11 van Den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

12 华琳, 阎岩, 张建. 关于对诊断一致性 Kappa 系统的探讨. 数理医药学杂志, 2006, 19(5): 518–520. [doi: 10.3969/j.issn.1004-4337.2006.05.033]

13 Cyr L, Francis K. Measures of clinical agreement for nominal and categorical data: The Kappa coefficient. Computers in Biology and Medicine, 1992, 22(4): 239–246. [doi: 10.1016/0010-4825(92)90063-S]

14 张嘉淏, 刘峰, 齐佳音. 一种基于 Bottleneck Transformer 的轻量级微表情识别架构. 计算机科学, 2022, 49(S1): 370–377.

15 王越, 王峰, 肖家赋, 等. 融合注意力机制和迁移学习的跨数据集微表情识别. 重庆理工大学学报 (自然科学), 2023, 37(1): 166–176.

16 Koch G, Zemel R, Salakhutdinov R. Siamese neural networks for one-shot image recognition. Proceedings of the 32nd International Conference on Machine Learning. ICML, 2015. 1–30.

17 Zhou L, Mao QR, Xue LY. Dual-inception network for cross-database micro-expression recognition. Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). Lille: IEEE, 2019. 1–5.

18 Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1867–1874.

19 Feng KJ, DeepLie: Detect lies with facial expression (computer vision) [Technical Report]. CS230. Stanford University, 2021.

(校对责编: 王欣欣)

© 中国科学院软件研究所 http://www.c-s-a.org.cn