



2024

Real-Time Multimodal Emotion Recognition

Presenter: 夏生杰

Date: 6/25/2024



目 录

CONTENTS

01

项目背景

02

项目目的

03

研究方法

04

研究过程

05

项目成果

06

参考文献





01

项目背景

Project Background

该项目选择探索文本、声音和视频输入，并开发一个集成模型，从所有这些来源收集信息，并以清晰可解释的方式显示。



概念定义



情感计算

情感计算 (*Affective Computing*) 是机器学习和计算机科学的一个领域，研究人类情感的识别和处理。

分类目标

Data Type	Categorical target
Textual	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
Sound	Happy, Sad, Angry, Fearful, Surprise, Neutral and Disgust
Video	

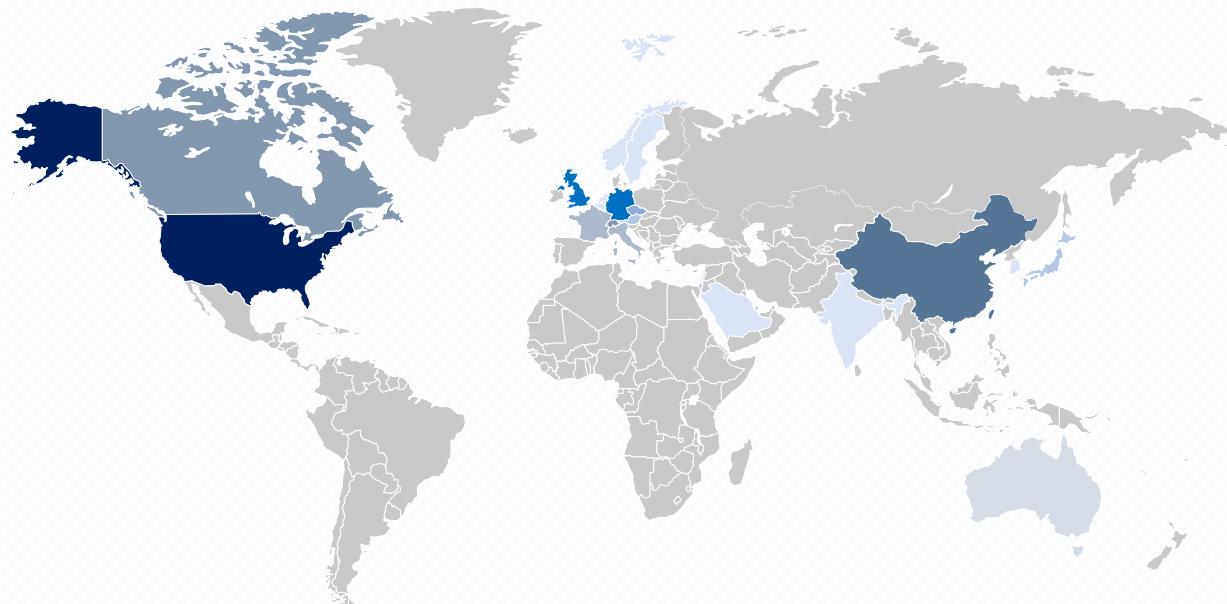
多模态情感识别

多模式情感识别 (*Multimodal Emotion Recognition*) 是一门相对较新的学科，旨在包括文本输入以及声音和视频。随着社交媒体的发展，这一领域一直在兴起，研究人员可以访问大量数据。最近的研究一直在探索潜在的指标来衡量来自不同渠道的情绪之间的一致性。





研究背景



自然语言处理 (*Natural Language Processing*)

通过文本进行情绪识别是一项具有挑战性的任务，它超越了传统的情绪分析：目标不是简单地从文本中检测中性、积极或消极的情绪，而是识别一组具有更高粒度特征的情绪。例如，愤怒或幸福之类的感觉可以包括在分类中。由于识别这种情绪甚至对人眼来说也可能很复杂，机器学习算法可能会获得好坏参半的性能。为了通过文本准确检测人类情绪，应该考虑到许多微妙之处，上下文依赖性是最关键的因素之一。

数字信号处理 (*Digital Signal Processing*)

语音情感识别的目的是从人的声音中自动识别出人的情感或身体状态。语音情感识别是基于离散情感分类系统的。在大多数情况下，文献只关注六种情绪标签，包括快乐、悲伤、愤怒、厌恶、恐惧和惊讶。语音情感识别的一般过程包括三个部分：信号处理、特征提取和分类。信号处理对原始音频信号应用声学滤波器，并将其分解为有意义的单元。特征提取是语音情感识别中的敏感点，因为特征既需要有效地表征人类语音的情感内容，又不依赖于词汇内容甚至说话者。最后，情感分类将特征矩阵映射到情感标签。

计算机视觉 (*Computer Vision*)

情绪识别等挑战通常无法通过经典的机器学习技术来解决。最近的研究论文都集中在几种深度学习技术上，其中包括人工神经网络(ANN)、卷积神经网络(CNN)、区域卷积神经网络(R-CNN)、快速区域卷积神经网络(Fast R-CNN)、递归神经网络(RNN)或长短期记忆(LSTM)。



02

项目目的

Project Purposes

该项目的目的是为求职者提供一个平台，通过声音和视频处理分析他们对一组预定义问题的答案，以及求职面试的非语言部分。



我们的目标是使用Tensorflow.js技术开发一个能够通过可视化用户界面提供实时情绪分析的模型。



项目目标



我们决定将两种类型的输入分开：
文本输入：例如平台上向某人提出的问题的答案。
视频输入：来自实时网络摄像头或存储在MP4或WAV文件中，我们从中分离音频和图像。

对于用于个性特征分类的文本数据，我们的主要目标是利用统计学习方法的使用，建立一个能够识别个人性格特征的工具，该文本包含了他对预先确定的个人问题的回答。我们选择将我们的性格特征检测模型应用于用户直接撰写的短文本。通过这种方式，我们可以轻松地针对特定的主题或问题，并提供要使用的语言水平的指示。因此，我们可以确保用于进行性格特征检测的文本数据与用于训练的数据一致，从而确保尽可能高的结果质量。

用于情绪识别的音频处理的主要目标是利用声音特征和语音特征分析来识别人的情绪状态，包括快乐、悲伤、愤怒等。通过分析音频中的声音特征、音调、语速等信息，可以准确识别出人的情绪状态，从而为情感智能系统提供更精准的情绪识别能力。

用于情绪识别的计算机视觉的主要目标是利用图像和视频中的面部表情、身体语言等视觉特征来识别人的情绪状态。通过分析人脸的表情特征、眼神、姿势等信息，可以准确识别出人的情绪状态，从而为情感智能系统提供更直观的情绪识别能力。





03

研究方法

Research Methodology

受制于计算机配置，我们选取了较为代表性的数据预处理方法与建模方法作为本项目的研究方法。

自然语言处理



数据预处理



预处理是我们将原始文本文档转换为经过清理的单词列表的过程。为了完成这个过程，我们首先需要对语料库进行标记化处理(Tokenize)。这意味着将句子分割为单词列表，也称为标记(Tokens)。其他预处理步骤包括使用正则表达式删除不需要的字符或重新格式化单词。例如，通常会将标记转换为小写，并删除一些对于理解文本并不重要的标点符号。删除停用词(Stopwords)以保留具有意义的单词也是一个重要步骤：这样可以摆脱像‘a’、‘the’或‘an’这样的常见单词。最后，有一些方法可以通过它们的语法根形式来替换单词：词干提取(Stemming)和词形还原(Lemmatization)的目标都是将一个词的各种派生形式归约为一个公共基本形式。具有相似含义的派生词族，如‘am’、‘are’、‘is’，将被替换为单词‘be’。最后，在词义消歧(Word Sense Disambiguation)的背景下，词性标注(Part-of-Speech tagging)用于标记语料库中的单词，指示其对应于特定词性的部分，基于其定义和上下文。这可以用于提高词形还原过程的准确性，或者只是更好地理解句子的含义。

标记化
句子分割为单词

正则表达式
小写转化，删除标
点符号、停用词

格式化单词
词干提取，词形还
原

词义消歧
词性标注

填充每个文档的标记序列以约束输入向量的形状。输入大小已固定为300：超过此索引的所有标记都将被删除。如果输入向量具有少于300个标记，则在向量的开头添加0，以便对形状进行归一化。填充序列的维度已经使用我们的训练数据的特征来确定。在任何预处理之前，每篇文章的平均字数为652个。在公式标准化、删除标点符号和停止语后，平均字数降至168个，标准差为68个。为了确保我们在分类中包含正确数量的单词，而不会丢弃太多信息，我们将填充维度设置为300，大致等于平均长度加上标准偏差的两倍。

数字信号处理



数据预处理 (cont.)



放大高频，平衡频谱



避免随时间丢失音频信号的频率轮廓



减少频谱能量泄漏



将信号从时间域转换到频率域

在开始特征提取之前，建议对音频信号应用预加重(Pre-emphasis)滤波器，以放大所有高频。预加重滤波器有几个优点：它可以平衡频谱，因为高频通常与低频相比具有较小的幅度，还可以避免傅里叶变换计算中的数值问题。在预加重滤波器之后，我们需要将音频信号分割成称为帧(Frame)的短期窗口。对于语音处理，窗口大小通常在20ms至50ms之间，两个连续窗口之间的重叠率在40%到50%之间。最常见的设置是帧大小为25ms，重叠为15ms(10ms窗口步长)。这一步的主要目的是避免随时间丢失音频信号的频率轮廓，因为音频信号本质上是非平稳的。实际上，信号中的频率特性随时间改变，因此在整个样本上应用离散傅里叶变换并没有太多意义。我们可以假设信号中的频率在很短的时间内是恒定的，因此可以在这些短时间窗口上应用离散傅里叶变换(Discrete Fourier Transform)，从而得到整个信号的频率轮廓的良好近似。在将信号分割成多个帧之后，我们将每个帧乘以汉明窗函数(Hamming window function)。这可以减少频谱泄漏或任何信号不连续，并提高信号的清晰度。例如，如果帧的开头和结尾不匹配，那么在离散傅里叶变换中会出现不连续情况。应用汉明函数可以确保开头和结尾匹配，同时平滑信号。离散傅里叶变换是数字信号处理各个领域中最广泛使用的变换，因为它可以将时域序列转换为频域。离散余弦变换(Discrete Cosine Transform)提供了音频信号频率内容分布的便捷表示。用于分析语音情感的大多数音频特征都是在频域中定义的，因为这更好地反映了音频信号的属性。

计算机视觉



数据预处理 (cont.)



计算机视觉的数据预处理方法有以下几种。图像尺寸调整(Image Resizing & Image Cropping): 图像的尺寸可能会因为不同的采集设备而存在差异，而且模型对于特定尺寸的图像有较好的适应性。因此，将图像调整为统一的尺寸是一种常见的数据预处理方法。图像标准化(Image Standardization): 将图像的像素值进行标准化，可以使得模型对图像的亮度和对比度变化不敏感。这种方法可以通过将像素值除以255(归一化)或者使用标准分数(z-score)标准化方法来实现。图像增强(Image Enhancement): 图像增强可以通过一系列的方法来提高图像的质量和信息含量，如旋转、翻转、剪裁、亮度/对比度调整等方法。图像平滑(Image Smoothing): 图像数据可能会受到噪音干扰，因此需要对图像进行去噪处理。数据增强(Data Augmentation): 数据增强是一种通过对原始数据进行变换生成新的训练数据的方法。这种方法可以有效地扩充训练集的规模和丰富数据的多样性，从而提高模型的泛化能力。特征提取(Feature Extraction): 在计算机视觉任务中，提取图像的特征信息是一种常见的预处理方法。通过提取图像的边缘、纹理、颜色等信息，可以减少输入数据的维度，从而降低模型的计算复杂度。

```
datagen = ImageDataGenerator(  
    zoom_range=0.2,      # randomly zoom into images  
    rotation_range=10,   # randomly rotate images  
    width_shift_range=0.1, # randomly shift images  
    horizontally=True,  
    height_shift_range=0.1, # randomly shift images vertically  
    horizontal_flip=True, # randomly flip images  
    vertical_flip=False) # randomly flip images
```

尺寸调整

- 裁剪(Cropping), Pixel values(48, 48)

标准化

- 归一化(Normalization)
- 标签编码(Label Encoding)

图像增强

- 灰色滤波器(gray filter)

图像平滑

- 均值滤波(Mean filtering)
- 中值滤波(Median filtering)
- 高斯滤波(Gaussian filter)

数据增强

- 放大(Zoom)
- 旋转(Rotate)
- 移动(Shift)
- 翻转(Flip)

特征提取

- 滑动窗口(Sliding window)
- 梯度方向直方图(Histogram of Oriented Gradients)
- 人脸特征点检测(Facial Landmark Detection)

自然语言处理



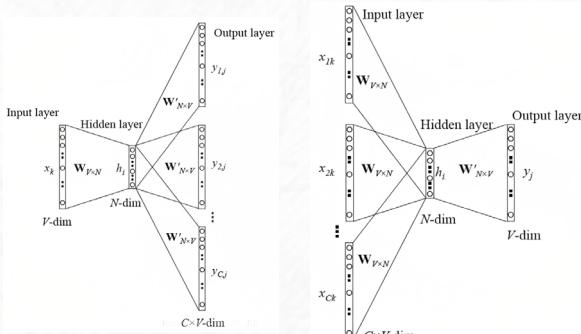
嵌入(Embedding)

- 词袋(Bag-of-Word)方法
 - TF-IDF
- 词嵌入(Word2Vec Embedding)
 - Skip Gram
 - Common Bag Of Words

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k \in d_j} n_{k,j}}$$

$$idf_i = \log \frac{D}{\{j: t_i \in d_j\} + 1}$$

$$TF-IDF = tf_{i,j} \times idf_i$$



Skip Gram训练过程:

- 将目标词进行独热表征作为模型的输入，其中词汇表的维度为V；
- 将输入词的独热向量乘以输入层到隐层的权重矩阵W得到隐藏层向量；
- 将隐藏层向量乘以隐藏层到输出层的权重矩阵W'得到V维的向量；
- 将计算得到的向量进行归一化指数函数激活处理得到V维的概率分布；
- 根据实际标签进行损失函数计算，使用梯度下降法对模型进行参数更新。
- 重复以上步骤，直到模型收敛为止。



建立模型



机器学习(Machine Learning)

- 多项式朴素贝叶斯(Multinomial Naive Bayes)
- 支持向量机(Support Vector Machine)

$$h(x) = \arg \max_k P(k) \prod_{i=1}^p p(x_i|k)$$

$P(k) = \frac{N_k + \alpha}{N + M\alpha}$ 条件独立(conditional independence)
多项式分布(multinomial distribution)

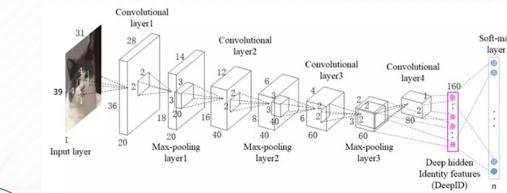
$P(x_i|k) = \frac{N_{kx_i} + \alpha}{N_k + M_i\alpha}$ 大数定律(law of large numbers)
拉普拉斯平滑(Laplace smoothing)

$$d = \frac{|w_1 * x_1 + w_2 * x_2 + \dots + w_n * x_n + b|}{\sqrt{w_1^2 + w_2^2 + \dots + w_n^2}} = \frac{|w^T * X + b|}{\|w\|}$$

$$\arg \max_{w,b} \min_{1 \leq i \leq n} \frac{y_i(w^T x_i + b)}{\|w\|}$$

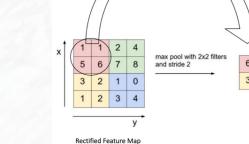
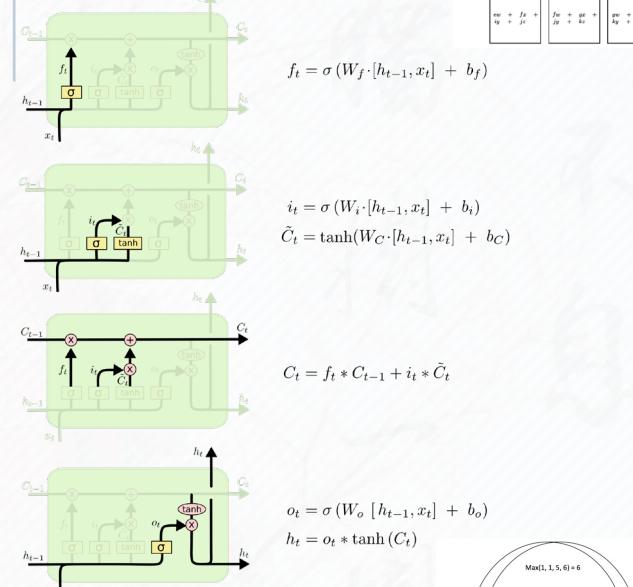
$\arg \min_{w,b,\xi} \frac{\|w\|^2}{2} + C \sum_{i=1}^n \xi_i$ 拉格朗日乘子(Lagrange multiplier)
原始-对偶(Primal-Dual)方法
松弛变量(Slack variable)

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, \dots, n, \xi_i \geq 0$$



深度学习(Deep Learning)

- 长短期记忆网络(LSTM)
- 卷积神经网络(CNN)



数字信号处理

深度学习 (Deep Learning)

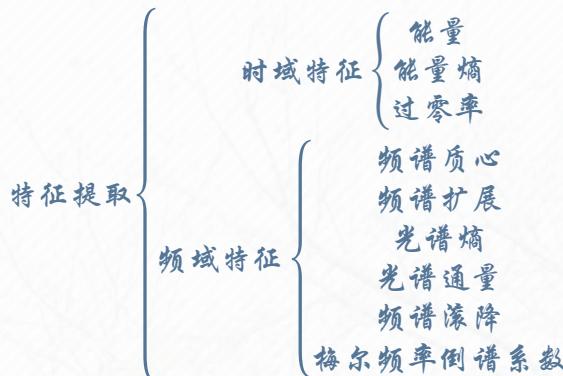
- 卷积神经网络
 - 批量归一化 (Batch Normalization)
- 长短期记忆神经网络

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1, \dots, x_m\}$
Parameters to be learned: γ, β
Output: $\{y_i = BN_{\gamma, \beta}(x_i)\}$

$$\mu_B \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{mini-batch mean}$$

$$\sigma_B^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2 \quad // \text{mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad // \text{normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma, \beta}(x_i) \quad // \text{scale and shift}$$


信号值的平方和，通过相应的帧长度进行归一化。(Energy)
 子帧归一化能量的熵，允许测量音频信号能量幅度的突变。(Entropy of energy)
 音频信号的符号变化率。(Zero Crossing rate)
 声音频谱的重心。(Spectral centroid)
 声音频谱的第二中心矩。(Spectral spread)
 子帧归一化光谱能量的熵。(Spectral entropy)
 两个连续帧的光谱归一化幅度之间的平方差，允许测量帧间的光谱变化。(Spectral flux)
 频谱90%的幅度分布集中的频率。(Spectral rolloff)
 将梅尔间隔滤波器组（一组三角形滤波器）应用于周期图，并计算每个滤波器中的能量。最后，我们对所有滤波器组能量的对数进行离散余弦变换，只保留前12个余弦变换。(Mel Frequency Cepstral Coefficients)

机器学习 (Machine Learning)

- 支持向量机 (Support Vector Machine)

建立模型

降维算法 (Dimensionality Reduction)

• 主成分分析 (Principal Component Analysis)

输入: n 维样本集 $D = (x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)})$, 要降维到的维数 n'
输出: 降维后的样本集 D'

- 对所有的样本进行中心化: $x^{(i)} \leftarrow x^{(i)} - \frac{1}{m} \sum_{j=1}^m x^{(j)}$
- 计算样本的协方差矩阵 XX^T
- 对矩阵 XX^T 进行特征值分解
- 取出最大的 n' 个特征值对应的特征向量 $(w_1, w_2, w_3, \dots, w_{n'})$, 将所有的特征向量标准化后, 组成特征向量矩阵 W
- 对样本集中的每一个样本 $x^{(i)}$, 转化为新的样本 $z^{(i)} = W^T x^{(i)}$
- 得到输出样本集 $D' = (z^{(1)}, z^{(2)}, z^{(3)}, \dots, z^{(m)})$

有时候, 我们不指定降维后的 n' 的值, 而是换种方式, 指定一个降维到的主成分比重阈值 t , 这个阈值 t 在 $(0, 1]$ 之间。假如我们的 n 个特征值为 $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_n$, 则 n' 可以通过下式得到:

$$\frac{\sum_{i=1}^{n'} \lambda_i}{\sum_{i=1}^n \lambda_i} \geq t$$

卡罗需-库恩-塔克条件 (Karush-Kuhn-Tucker Conditions)

$$f(x) = W^T \phi(x) + b$$

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)$$

$$s.t. \sum_{i=1}^m \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, m$$

线性核函数: $k(x_i, x_j) = x_i^T x_j$

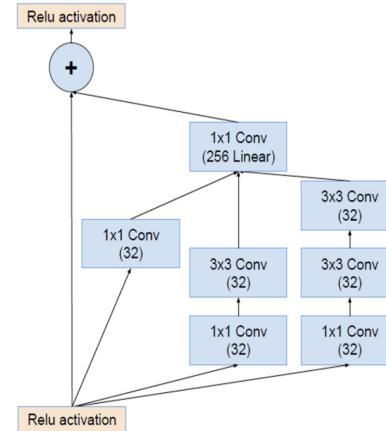
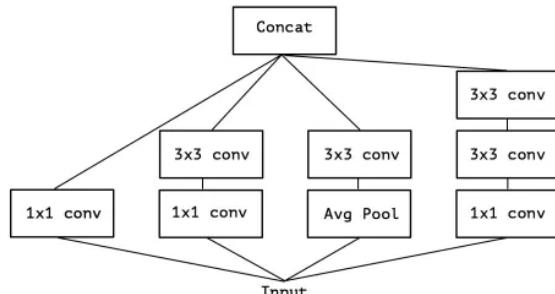
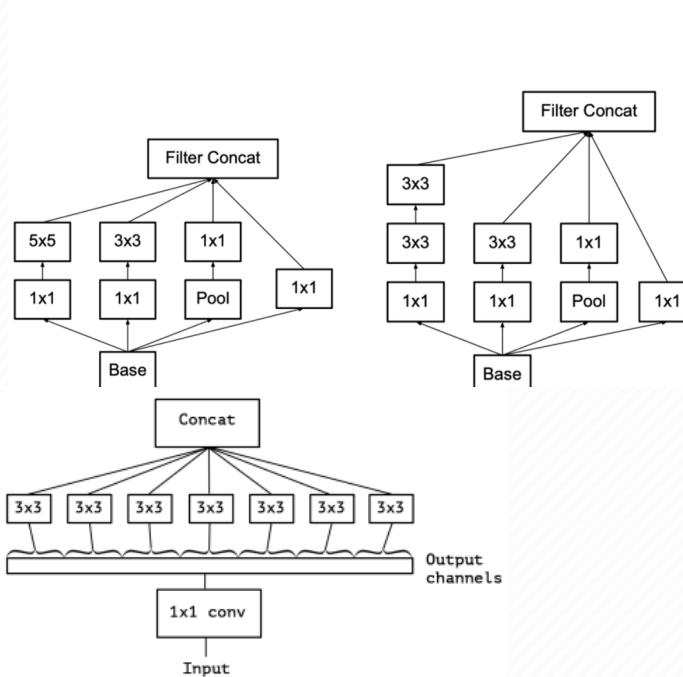
多项式核函数: $k(x_i, x_j) = (x_i^T x_j)^d$

高斯核函数: $k(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$

计算机视觉

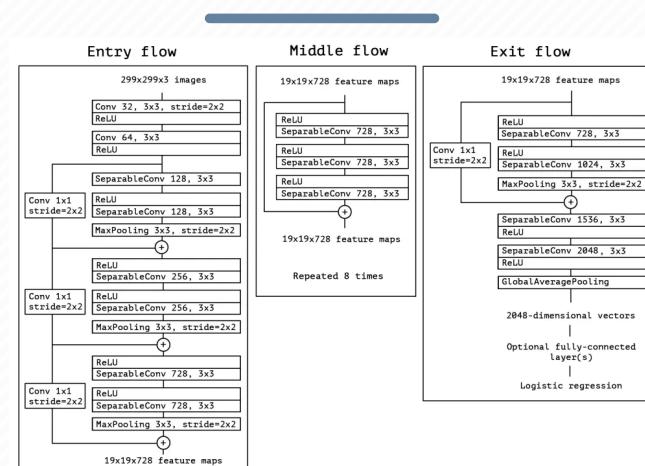


建立模型



Inception: 核心思想就是将不同的卷积层通过并联的方式结合在一起，经过不同卷积层处理的结果矩阵在深度这个维度拼接起来，形成一个更深的矩阵。

Xception: 采用深度可分离卷积(Depthwise Separable Convolution)块来替换原来 Inception v3 中的多尺寸卷积核特征响应操作。



支持向量机

卷积神经网络



04

研究过程

Research Process

文本、声音和视频输入的模型训练过程及结果展示。

01

自然语言处理

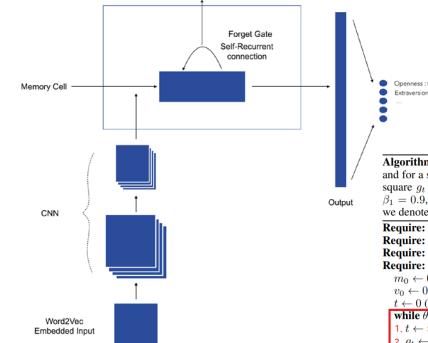
我们使用的是项研究[1]中收集的数据。它包括34名心理学学生（29名女性和5名男性，年龄从18岁到67岁，平均26.4岁）每天提交的2468份书面材料。提交的书面材料是以未评级的课程作业的形式提交的。对于每项作业，学生每天至少要就特定主题写20分钟的文章。这些数据是在1993年至1996年为期两周的夏季课程中收集的。每个学生连续10天完成每日写作。学生的个性得分是通过回答五大性格特质（BFI）[2]来评估的。五大性格特质是一份44项的自我报告问卷，为五种人格特征中的每一种提供分数。每个项目都由短语组成，并使用5分制进行评分，范围从1（强烈反对）到5（强烈同意）。数据源中的一个实例包括一个ID、实际文章和五大人格特征的五个分类标签。标签最初的形式是“是”（“y”）或“否”（“n”），表示给定特质的得分高或低。需要注意的是，分类标签是根据一份相当短的自我报告问卷的答案应用的：与心理特征的复杂性相比，该测试相对简单，而且认知偏见阻碍了用户对自己的特征进行完全准确的评估，因此数据中可能存在不可忽略的偏差。

我们选择了一种基于一维卷积神经网络和递归神经网络的神经网络架构。一维卷积层的作用与特征提取相当：它允许在文本数据中找到模式。然后使用长短期记忆单元来利用自然语言的序列性质：与假设输入彼此独立的常规神经网络不同，这些架构通过序列逐渐积累和捕获信息。长短期记忆（LSTM）具有长时间选择性记忆模式的特性。我们的最终模型首先包括由以下四层组成的3个连续块：一维卷积层-最大池化-空间丢弃-批处理规范化。每个块的卷积滤波器的数量分别为128、256和512，内核大小为8，最大池大小为2，丢弃率为0.3。在这三个块之后，我们选择堆叠3个记忆（LSTM）单元，每个单元有180个输出。最后，在最后一个分类层之前添加128个节点的完全连接层。



支持向量机

通过随机梯度下降手动实现了基于文本分类任务的线性支持向量机模型。



Algorithm 1: Adam, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation, \hat{g}_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t .

```

Require:  $\alpha$ : Stepsize
Require:  $\beta_1, \beta_2 \in [0, 1]$ : Exponential decay rates for the moment estimates
Require:  $f(\theta)$ : Stochastic objective function with parameters  $\theta$ 
Require:  $\theta_t$ : Current parameter vector
 $m_0 \leftarrow 0$  (Initialize 1st moment vector)
 $v_0 \leftarrow 0$  (Initialize 2nd moment vector)
 $t \leftarrow 0$  (Initialize timestep)
while  $\theta_t$  not converged do
    1.  $t \leftarrow t + 1$ 
    2.  $g_t \leftarrow -\nabla_{\theta} f_t(\theta_{t-1})$  (Get gradients w.r.t. stochastic objective at timestep  $t$ )
    3.  $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$  (Update biased first moment estimate)
    4.  $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$  (Update biased second raw moment estimate)
    5.  $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$  (Compute bias-corrected first moment estimate)
    6.  $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$  (Compute bias-corrected second raw moment estimate)
    7.  $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$  (Update parameters)
end while
return  $\theta_t$  (Resulting parameters)

```

神经网络

使用卷积神经网络与长短期记忆神经网络相结合的方式，损失函数为类别交叉熵（Categorical Cross Entropy），优化器为自适应矩估计（Adaptive Moment Estimation），批次训练大小为32，训练轮数为135轮。

[1] Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: language use as an individual difference. Journal of personality and social psychology, 77(6), 1296.

[2] John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. Journal of personality and social psychology.



RAVDESS								
Emotions	Happy	Sad	Angry	Scared	Disgusted	Surprised	Neutral	Total
Man	96	96	96	96	96	96	96	672
Woman	96	96	96	96	96	96	96	672
Total	192	192	192	192	192	192	192	1344

数字信号处理

我们使用的数据集为该研究中^[1]提出的数据集包含24名专业演员（12名女性，12名男性），用中性的北美口音说出两个词汇匹配的语句。言语包括平静、快乐、悲伤、愤怒、恐惧、惊讶和厌恶的表情，歌曲包括平静、高兴、悲伤、生气和恐惧的情绪。每个表情都是在两个情绪强度水平上产生的（正常、强烈），还有一个额外的中性表情。为了限制训练阶段的过度拟合，我们将数据集分为训练集（80%）和验证集（15%）测试集（5%）。

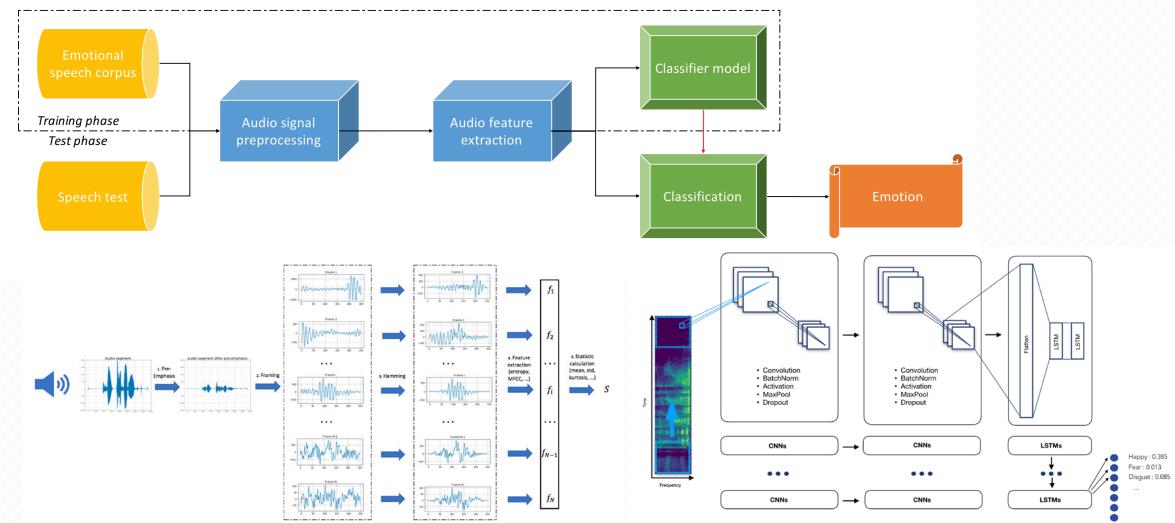
02

支持向量机

语音情感识别的经典方法包括对音频信号应用一系列滤波器，并将其划分为几个窗口（固定大小和时间步长）。然后，针对每个帧提取时域（过零率、能量和能量熵）和频域（谱熵、质心、扩散、通量、滚降和梅尔频率倒谱系数）的特征。然后，我们计算这些特征中的每一个的一阶差分，以捕捉信号中的逐帧变化。最后，我们计算这些特征的以下全局统计信息：均值、中值、标准差、峰度、偏度、1%百分位、99%百分位数、最小值、最大值和范围，从而为每个音频信号获得360个候选特征的向量，并用核函数训练一个简单的SVM分类器来预测语音中检测到的情绪。

神经网络

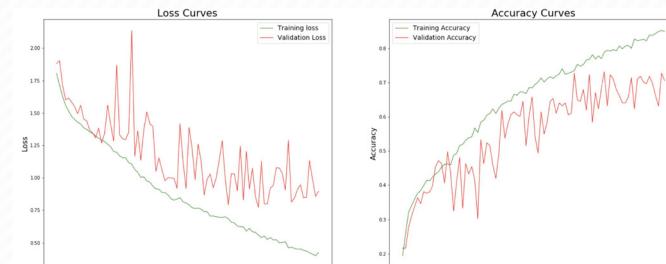
时间分布卷积神经网络的主要思想是在对数梅尔谱图上应用滚动窗口（固定大小和时间步长）。这些窗口中的每个窗口将是由四个局部特征学习(LFLB)组成的卷积神经网络的入口，并且这些卷积网络中的每个的输出将被馈送到由2个单元长短期记忆(LSTM)组成的递归神经网络中，以学习长期上下文依赖性。最后，使用具有归一化指数(softmax)激活的完全连接层来预测在语音中检测到的情绪。损失函数为类别交叉熵(Categorical Cross Entropy)，优化器为随机梯度下降(Stochastic Gradient Descent)，批次训练大小为64，训练轮数为100轮。



PCA dimension	linear	poly (2)	poly (3)	rbf
None	51.67%	54.28%	52.79%	56.51%
140	53.53%	50.19%	52.79%	59.48%
120	55.02%	50.56%	52.79%	58.74%
100	52.79%	48.33%	52.79%	58.36%

$$g = \nabla_{\theta} L(x^{(i)}, y^{(i)}, \theta)$$

Name	Output dim	Kernel	Stride	Other
LFLB ₁	Conv M × N × 64	3 × 3	1 × 1	64 filters
	BatchNorm M × N × 64	—	—	—
	Activation M × N × 64	—	—	Elu
	Max Pool M/2 × N/2 × 64	2 × 2	2 × 2	—
LFLB ₂	Dropout M/2 × N/2 × 64	—	—	0.3
	Conv M/2 × N/2 × 64	3 × 3	1 × 1	64 filters
	BatchNorm M/2 × N/2 × 64	—	—	—
	Activation M/2 × N/2 × 64	—	—	Elu
LFLB ₃	Max Pool M/8 × N/8 × 64	4 × 4	4 × 4	—
	Dropout M/8 × N/8 × 64	—	—	0.3
	Conv M/8 × N/8 × 128	3 × 3	1 × 1	128 filters
	BatchNorm M/8 × N/8 × 128	—	—	—
LFLB ₄	Activation M/8 × N/8 × 128	—	—	Elu
	Max Pool M/32 × N/32 × 128	4 × 4	4 × 4	—
	Dropout M/32 × N/32 × 128	—	—	0.3
	Conv M/32 × N/32 × 128	3 × 3	1 × 1	128 filters
LSTM ₁	BatchNorm M/32 × N/32 × 128	—	—	—
	Activation M/32 × N/32 × 128	—	—	Elu
	Max Pool M/128 × N/128 × 128	4 × 4	4 × 4	—
	Dropout M/128 × N/128 × 128	—	—	0.3
LSTM ₂	256	—	—	—
Fully Connected	7 labels	—	—	—



最高精度: 74%

[1] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.

03



计算机视觉

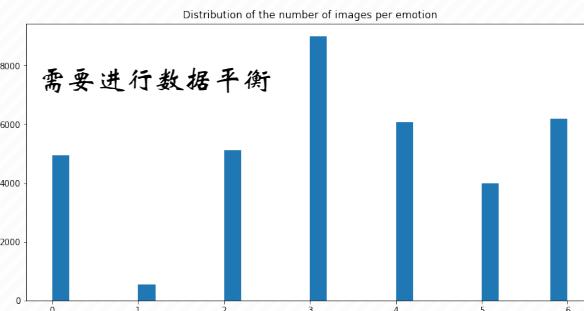


我们使用的数据集[1]数据由 48×48 像素的人脸灰度图像组成。人脸已经被自动配准，使得人脸或多或少地居中，并且在每个图像中占据大约相同的空间量。任务是根据面部表情中显示的情绪将每张脸分为七类（0=愤怒，1=厌恶，2=恐惧，3=快乐，4=悲伤，5=惊讶，6=中性）之一。用于训练的实例(instance)数为28709，测试的实例数为3589。



支持向量机

使用高斯核函数

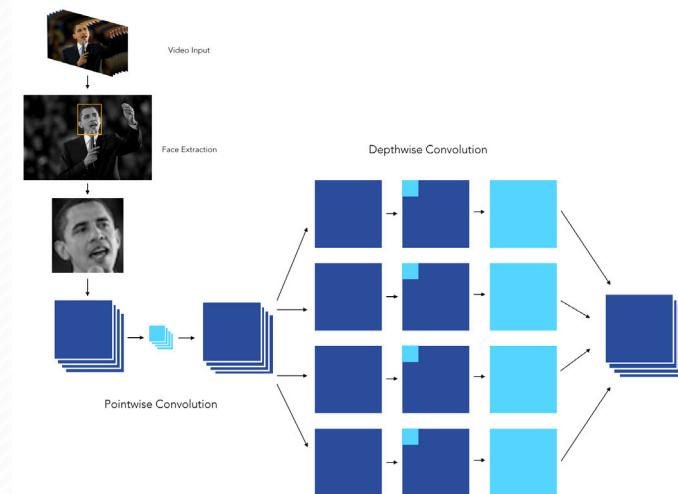


损失函数为类别交叉熵(Categorical Cross Entropy)，优化器为自适应矩估计(Adaptive Moment Estimation)，批次训练大小为128，训练轮数为150轮。

我们可以绘制类激活图(class activation maps)，它显示了已被最后一个卷积层激活的像素。我们注意到像素是如何被不同地激活的，这取决于被标记的情绪。快乐似乎取决于与眼睛和嘴巴相连的像素，而悲伤或愤怒似乎更多地与眉毛有关。



神经网络





05

项目成果

Project Results

通过进行该项目获得的收获。

>>> **Flask**

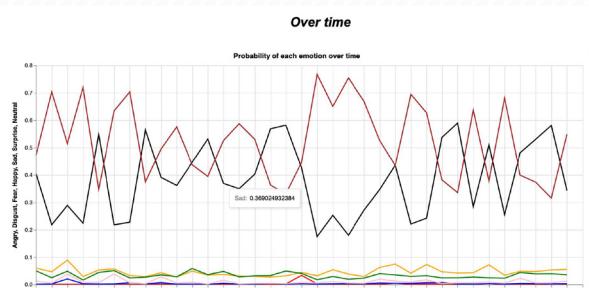
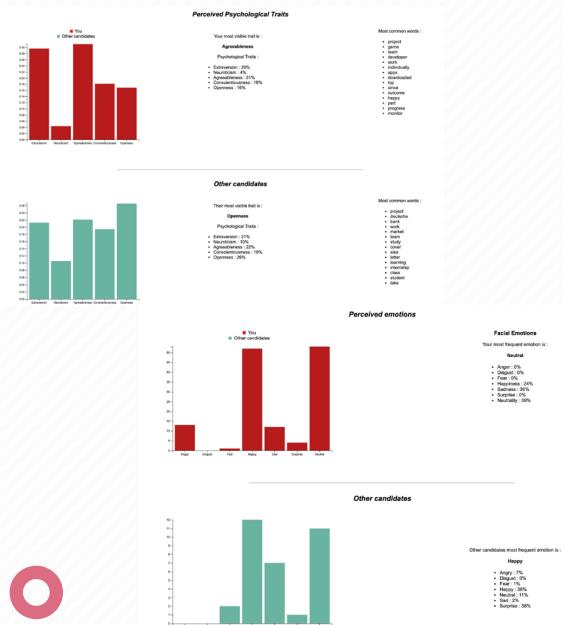
Flask是一个基于Python开发并且依赖jinja2模板和Werkzeug WSGI服务的一个微型框架，对于Werkzeug本质是Socket服务端，其用于接收http请求并对请求进行预处理，然后触发Flask框架，开发人员基于Flask框架提供的功能对请求进行相应的处理，并返回给用户，如果要返回给用户复杂的内容时，需要借助jinja2模板来实现对模板的处理，即：将模板和数据进行渲染，将渲染后的字符串返回给用户浏览器。

为了实现我们的模型，我们选择创建一个开源网络应用程序。该平台的目的是以直观且易于访问的方式向我们所有的情绪识别模型提供信息。它允许用户基于对直接通过平台发送的视频、音频或文本摘录的分析，实时获得对其情绪或性格特征的个性化评估。该工具可以用于希望练习面试技能的求职者：求职者可以尽可能多地训练自己来回答问题，每次都可以获得我们的算法感知到的情绪/个性特征的摘要，并与其他求职者进行比较。

每个通信通道（音频、视频、文本）都有一个专用页面允许对用户进行评估。每一页都会问一个典型的面试问题，例如：“告诉我们你上一次展现领导力是什么时候”。音频/视频摘录（通过计算机麦克风/网络摄像头录制）或文本块保存后即可检索，并由我们的算法处理（在文本频道的情况下，用户还可以上传一个.pdf文档，该文档将由我们的工具解析）。

一旦用户记录或键入了他的答案，他就会被重定向到总结页面。例如，在视频面试的情况下，这种评估不仅可以让他知道自己在我们的模型识别的每种情绪中的“分数”，还可以让他了解其他候选人的平均分数：这样他就可以重新定位自己，并随意调整态度。我们认为，在分析中包括一种基准有助于用户意识到他或她相对于普通候选人的地位。

文本和视频/音频摘要略有不同：对于文本面试摘要，我们不仅选择显示用户和其他候选人的已识别性格特征的百分比分数，还选择显示答案中最常用的单词。对于视频和音频面试摘要，我们显示了用户和其他候选人的感知情绪得分。



Video Interview

Tell us about the last time you showed leadership.

Start Recording

You will have 45 seconds to discuss the topic mentioned above. Due to restrictions, we are not able to redirect you once the video is over. Please move your URL to /video_dash instead of /video_1 once over. You will be able to see your results then.

How does it work ?

Back

Interview Simulator

Video Interview

Use the video interview simulator and get a feedback on how our algorithm interprets your facial emotions compared to other candidates.

You will be provided a feedback on your facial emotions such as :

- Anger
- Happiness
- Surprise
- Sadness
- Surprise
- Disgust

Audio Interview

Use the audio interview simulator and get a feedback on how our algorithm interprets your vocal emotions compared to other candidates.

You will be provided a feedback on your vocal emotions such as :

- Anger
- Happiness
- Fear
- Sadness
- Surprise
- Disgust

Text Interview

Use the text interview simulator and get a feedback on how our algorithm interprets your psychological traits compared to other candidates.

You will be provided a feedback on your Big Five Psychological traits, which include :

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

Video Interview **Audio Interview** **Text Interview**

Audio Interview

Tell us about the last time you showed leadership.

Start Recording

After pressing the button above, you will have 15sec to answer the question.

How does it work ?

Back

Number of Faces : 1

Emotional report : Face #1

Angry	: 0.053
Disgust	: 0.0
Fear	: 0.053
Happy	: 0.001
Sad	: 0.153
Surprise	: 0.002
Neutral	: 0.738

Text Interview

Tell us about the last time you showed leadership.

Or upload your Cover Letter :

Choisir un fichier / Auswahl ficher choisir

Start Analysis





06

参考文献

References

项目参考的资源。



- [1] Ververidis, D., & Kotropoulos, C. (2003). A Review of Emotional Speech Databases.
- [2] Vogt, T., André, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, 75-91.
- [3] Bhakre, S. K., & Bang, A. (2016, September). Emotion recognition on the basis of audio signal using Naive Bayes classifier. In *2016 International conference on advances in computing, communications and informatics (ICACCI)* (pp. 2363-2367). IEEE.
- [4] Ke, X., Zhu, Y., Wen, L., & Zhang, W. (2018). Speech emotion recognition based on SVM and ANN. *International Journal of Machine Learning and Computing*, 8(3), 198-202.
- [5] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135). IEEE.
- [6] Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9), 1162-1181.
- [7] Burgos, W. (2014). Gammatone and MFCC features in speaker recognition.
- [8] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2), 190-202.
- [9] Wanare, M. A. P., & Dandare, S. N. (2014). Human emotion recognition from speech. system, 6.
- [10] Vogt, T., & André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation.
- [11] Giannakopoulos, T., & Pikrakis, A. (2014). Introduction to audio analysis: a MATLAB® approach. Academic Press.
- [12] Basharirad, B., & Moradhaseli, M. (2017, October). Speech emotion recognition methods: A literature review. In *AIP conference proceedings* (Vol. 1891, No. 1). AIP Publishing.
- [13] Wang, Z., & Xue, X. (2014). Multi-class support vector machine. *Support vector machines applications*, 23-48.
- [14] Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *Plos one*, 10(12), e0144610.
- [15] Vogt, T. (2010). Real-time automatic emotion recognition from speech.
- [16] Aytar, Y., Vondrick, C., & Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29.
- [17] Casale, S., Russo, A., Sciebba, G., & Serrano, S. (2008, August). Speech emotion classification using machine learning algorithms. In *2008 IEEE international conference on semantic computing* (pp. 158-165). IEEE.
- [18] Chen, L., Mao, X., Xia, Y., & Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital signal processing*, 22(6), 1154-1160.
- [19] Vaishnav, S., & Mitra, S. (2016). Speech emotion recognition: a review. *International Research Journal of Engineering and Technology (IRJET)*, 3(04).
- [20] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A. (2018). Speech Emotion Recognition: Methods and Cases Study. *ICAART* (2), 20.
- [21] Shahsavaran, S. (2018). Speech emotion recognition using convolutional neural networks.
- [22] Seehapoch, T., & Wongthanavasu, S. (2013, January). Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)* (pp. 86-91). IEEE.
- [23] Chavhan, Y., Dhore, M. L., & Yesaware, P. (2010). Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, 1(20), 6-9.
- [24] El Ayadi, M., Kamel, M. S., & Karay, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3), 572-587.
- [25] Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision support systems*, 115, 24-35.
- [26] Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., & Dehak, N. (2018, September). Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts. In *Interspeech* (pp. 247-251).
- [29] Tighe, E. P., Ureta, J. C., Pollo, B. A. L., Cheng, C. K., & de Dios Bulos, R. (2016, July). Personality Trait Classification of Essays with the Application of Feature Reduction. In *SAAIP@ IJCAI* (pp. 22-28).
- [30] Tarnowski, P., Kołodziej, M., Majkowski, A., & Rak, R. J. (2017). Emotion recognition using facial expressions. *Procedia Computer Science*, 108, 1175-1184.
- [31] Karhunen, J., Raiko, T., & Cho, K. (2015). Unsupervised deep learning: A short review. *Advances in independent component analysis and learning machines*, 125-142.
- [32] Duncan, D., Shine, G., & English, C. (2016). Facial emotion recognition in real time. *Computer Science*, 1-7.
- [33] Pramerdorfer, C., & Kampel, M. (2016). Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*.
- [34] Shanmugamani, R. (2018). Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras. Packt Publishing Ltd.
- [35] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
- [36] Kim, M. H., Joo, Y. H., & Park, J. B. (2005). Emotion detection algorithm using frontal face image. *제어로봇시스템학회: 학술대회논문집*, 2373-2378.
- [37] Ruiz-Garcia, A., Elshaw, M., Altahhan, A., & Palade, V. (2018). A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. *Neural Computing and Applications*, 29, 359-373.
- [38] Arriaga, O., Valdenegro-Toro, M., & Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

