

引用格式:薛艳飞,张建明.基于对抗训练的跨语料库语音情感识别方法[J].微电子学与计算机,2021,38(3):77-83. [XUE Y F, ZHANG J M. Cross-corpus speech emotion recognition based on adversarial training[J]. Microelectronics & Computer, 2021,38(3):77-83.]

基于对抗训练的跨语料库语音情感识别方法

薛艳飞,张建明

(江苏大学 计算机科学与通信工程学院,江苏 镇江 212013)

摘 要:在跨语料库语音情感识别中,训练和测试数据分布的差异变得非常明显,导致验证和测试性能差别很大.针对该问题,提出一种基于对抗训练的跨语料库语音情感识别方法.该方法通过语料库之间的对抗训练能有效地缩小不同语料库之间的差异,提升模型对域不变情感特征的提取能力.同时,通过引入多头自注意力机制,对语音序列中不同位置元素之间的依赖关系进行序列建模,增强序列中情感显著特征的提取能力.在以 IEMOCAP 为源域、MSP-IMPRO 为目标域和在以 MSP-IMPRO 为源域、IEMOCAP 为目标域上的实验表明,所提出方法的相对 UAR 性能相比于基准方法分别提升了 0.91%~12.22%和 2.27%~6.90%.因此,在目标域标注缺失的情况下,所提出的跨语料库语音情感识别方法具有更好的域不变情感显著特征的提取能力.

关键词:领域适应;对抗训练;卷积神经网络;循环神经网络;自注意力机制

中图分类号: TP183

文献标识码: A

文章编号: 1000-7180(2021)03-0077-07

Cross-corpus speech emotion recognition based on adversarial training

XUE Yan-fei, ZHANG Jian-ming

(School of Computer Science and Communication Engineering, Jiangsu University,
Zhenjiang 212013, China)

Abstract: The difference in data distributions becomes very clear when the training and testing data come from different corpora, causing a large performance gap between validation and testing performance. To solve this problem, a cross-corpus speech emotion recognition method based on adversarial training is proposed. The proposed method can effectively eliminate the differences between different corpora with the adversarial training of corpora, and improve the extracting ability of domain-invariant emotion features. At the same time, model the relative dependence of different position elements in the speech sequence to enhance the emotion-salient features extracting ability of the sequence by introducing the multi-head attention mechanism. When the experiment applies IEMOCAP as the source domain and MSP-IMPRO as the target domain, the results are superior to the benchmark methods about 0.91%~12.22%. Meanwhile, the experiment applies MSP-IMPRO as the source domain and IEMOCAP as the target domain, the results also achieve better performance than the benchmark methods about 2.27%~6.90%. Therefore, in the case of the absence of emotion labels of the target domain, the proposed cross-corpus speech emotion recognition method is more beneficial to extracting domain-invariant emotion salient features.

Key words: domain adaptation; adversarial training; convolutional neural network, recurrent neural network; self-attention mechanism

1 引言

跨语料库语音情感识别存在的问题是许多实际应用中的关键问题. 当训练和测试数据属于不同的语料库时会导致训练和测试之间存在显著的性能差异. 随着深度领域适应技术的快速发展, 越来越多的研究者开始关注跨语料库语音情感识别的研究. 如何缓解不同语料库边缘概率分布不一致导致的跨语料库语音情感识别性能不足的问题是当前面临的一个重要挑战.

Liu 等^[1] 针对跨语料库语音情感识别问题提出了一种基于领域自适应子空间学习的投影矩阵学习方法 (DoSL), 将源语音信号和目标语音信号从原始特征空间转换到具有相似特征的标签空间, 在源语料库上学习的分类器可以有效地预测未标注的目标域语音的情感状态. Mao 等^[2] 提出了一种改进的情感判别和域不变的特征学习模型 (E_EDFLM). 通过一个简单的前向神经网络模型, 可以得到输入数据的一个层次非线性的高层特征表示. 同时, 对输入数据的高层特征表示分别添加情感标签预测器和域标签预测器两个分类器使学到的特征具有情感判别性和域不变性. 进一步地, 将输入数据分解成情感相关特征和情感无关特征两部分以在特征学习过程中排除一些情感无关的因素. Abdelwahab 等^[3] 提出了一种基于域对抗训练的声学情感识别方法, 研究探讨了如何使用对抗性多任务训练来提取训练域和测试域之间的公共表征. 在 Arousal、Valence 和 Dominance 情感属性分类任务上验证了使用无标记训练数据进行对抗训练有利于情感识别. Gideon 等^[4] 提出了一种对抗判别域泛化方法 (ADDog), 该方法通过迭代的方式来拉近源域和目标域数据集的情感特征分布, 从而提升跨语料库情感泛化的能力. 试验结果表明, 即使在没有标注的目标域数据可用的情况下, ADDog 也能一致性收敛并生成泛化性更好的表征. Neumann 等^[5] 提出了一种跨语言和多语言的语音情感识别模型, 通过训练了一个基于注意力的卷积神经网络 (ACNN) 来对语音情感属性中的 Arousal 和 Valence 任务进行识别. Lee 等^[6] 通过对两种高度异质化语言语音情感的研究, 提出了一种基于深度神经网络 (DNN) 的多语言和跨语言语音情感识别方法.

虽然基于深度学习的语音情感识别方法在跨语料库语音情感识别中已经取得了初步进展, 但是如何学习得到语料库不变的共享情感特征空间分布,

缓解不同语料库语音情感特征边缘概率分布不一致导致的模型性能下降问题仍是当前亟待解决的问题. 针对该问题, 本文提出了一种基于对抗训练的跨语料库语音情感识别方法, 主要贡献和创新点概括如下:

(1) 提出一种基于语料库对抗训练的卷积循环多头自注意力网络. 在该方法中, 通过不同语料库之间对抗训练的方式提升模型对语料库无关情感特征的建模能力, 以此来缓解训练和测试数据边缘概率不匹配所造成的模型性能大幅下降的问题.

(2) 通过引入多头自注意力机制, 可以对语音序列中不同位置的元素之间的相对依赖关系进行时间序列建模, 从而增强模型对序列中情感显著部分的特征提取能力, 得到源域和目标域共享的情感显著空间.

2 语料库不变情感特征学习

在本文中, 为了解决跨语料库语音情感识别中的挑战, 提出一种基于对抗训练的卷积循环多头自注意力网络 (Convolutional Recurrent Multi-Head Self-Attention Network Based on Adversarial Training, CRMHSAN_AT), 用于跨语料库语音情感识别, 其结构如图 1 所示. CRMHSAN_AT 主要包括输入层 I、域不变情感特征编码器 G、情感分类器 F 和域判别器 D 四部分. 其中, 输入层 I 以特征元素之间具有高度相关性的 FBanks 声学特征作为原始输入特征, 用于域不变情感特征的学习; 域不变情感特征编码器 G 用于对语料库无关情感特征的提取; 情感分类器 F 用于对域不变的情感语义特征进行情感分类; 域判别器 D 通过对抗训练达到域混淆的目的, 促进域不变情感特征的学习.

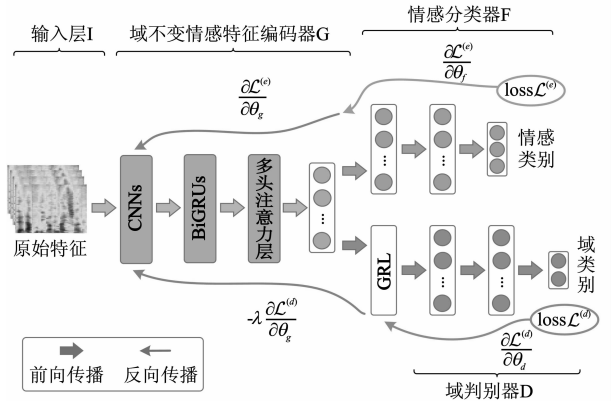


图 1 CRMHSAN_AT 网络结构

2.1 输入层

在语音情感识别领域,FBanks 特征相邻特征之间具有高度的相关性,相比于 MFCCs 特征包含更多的声学信息且更适宜神经网络的学习,是目前主流的深度语音情感识别模型输入特征之一.因此,该特征被选作 CRMHSAN_AT 的输入特征.

从原始的语音信号中提取 FBanks 特征,首先要对语音信号进行预处理,预处理包含了预加重、分帧、加窗等.由于语音信号是一个时变信号,对语音信号的分析都是在语音信号的短时平稳性假设条件下进行的.因此,需要对长语音进行分帧操作,CRMHSAN_AT 模型中取 25 ms 为一帧;为了保持帧与帧之间的连贯性,相邻帧之间还需要取 15 ms 的帧重叠;然后,对分帧后的语音进行加窗,用以对帧的端点做频谱平滑处理,缓解频谱泄漏.以语音 Y 为例,分帧加窗后即可得到 N 帧信号,其形式化表达如公式(1)所示.

$$Y = [Y(1), Y(2), \dots, Y(N)] \tag{1}$$

然后对每帧信号进行 FBanks 特征提取. FBanks 特征描述了人耳频率的非线性特性. 它的提取过程分为以下几个步骤:快速傅里叶变换、三角带通滤波和对数能量计算. 经过以上处理便可得到 t 帧信号对应的 FBanks 特征序列 X ,以此作为 CRMHSAN_AT 模型的输入特征,如公式(2)所示.

$$X = [X(1), X(2), \dots, X(N)] \tag{2}$$

2.2 域不变情感特征编码器

域不变情感特征编码器的作用是对语料库不变的情感特征进行编码,学习得到源域和目标域不变的共享情感特征空间,以此提升模型对目标语料库情感特征空间的泛化能力. 域不变情感特征编码器由两层卷积神经网络(CNN)层、两层双向门控循环单元(GRU)层、多头自注意力(Multi-Head Self-Attention)层和一层全连接(Full Connection, FC)层构成. 其中,卷积神经网络层结构如表 1 所示.

表 1 卷积层体系结构

类型	输出维度	核尺寸	步长	填充
卷积	40×751×128	7×7	1×1	SAME
批归一化	40×751×128	—	—	—
非线性激活	40×751×128	—	—	—
最大池化	20×376×128	2×2	2×2	SAME
卷积	1×370×128	20×7	1×1	VALID
批归一化	1×370×128	—	—	—
非线性激活	1×370×128	—	—	—
最大池化	1×74×128	1×5	1×5	SAME

在卷积层提取的高层语义特征的基础之上将其重朔为 74×128 维之后送入双向门控循环单元进行上下文依赖的情感特征学习. GRU 是对 RNN 的改进,将 LSTM 中的“输入门”和“遗忘门”整合为“更新门”,同时配置一个“重置门. 对于卷积层提取的中间层特征序列 $M=[m_1, m_2, \dots, m_t, \dots, m_T]$,GRU 的计算过程分别如公式(3)~(6)所示:.

$$z_t = \sigma(W_z \cdot [h_{t-1}, m_t]) \tag{3}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, m_t]) \tag{4}$$

$$\tilde{h}_t = \tanh(W_t \cdot [r_t \odot h_{t-1}, m_t]) \tag{5}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{6}$$

式中, σ 为 sigmoid 激活函数, z_t 为更新门, r_t 为重置们, W_t, W_r 分别为 z_t 和 r_t 对应层的神经元参数, h_{t-1} 为 $t-1$ 时刻隐藏层的输出, $\tilde{h}_t$ 为 t 时刻候选隐含层, h_t 为 t 时间步隐藏层输出.

为了得到语音序列中上下文相关的情感信息,CRMHSAN_AT 方法中采用双向门控循环单元(Bidirectional Gated Recurrent Units, BiGRU)抽取源域和目标域参数共享的高层语义表示,分别使用 $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 表示. 同时,使用非线性变换可以得到最终的隐向量 $h_t = \sigma([\vec{h}_t, \overleftarrow{h}_t])$. 因此,BiGRU 的输出 H 形式化表示如公式(7) 所示.

$$H = [h_1, h_2, \dots, h_t, \dots, h_T] \tag{7}$$

式中, $H \in \mathbb{R}^{T \times d}$, T 为序列长度, d 为特征维度. 在 CRMHSAN_AT 方法中,使用两层 BiGRU 作为时序特征提取器,每层 GRU 的节点数为 128,将得到的所有时间步的输出特征作为多头注意力层的输入特征.

多头自注意力机制是 Vaswani 等人在文献[7]的 Transformer 网络中首次引入,是由自注意力机制发展而来的. CRMHSAN_AT 方法中引入多头注意力的目的是为了对序列中不同位置的元素之间的相对依赖关系进行建模,增强从语音序列中抽取情感显著特征的能力. 对于 BiGRU 提取的情感语义特征 $H = [h_1, h_2, \dots, h_t, \dots, h_T]$,多头自注意力计算流程如图 2 所示.

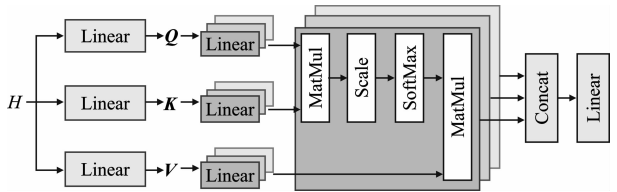


图 2 多头自注意力机制的并行结构

自注意力机制对输入序列的每个元素都可以通

过一个线性变换投影到三个不同的表示形式: query、key 和 value. 具体说来, 给定情感语义特征序列 H , 对应的线性变换方式如公式(8) 所示.

$$q_t = w_q^T h_t, k_t = w_k^T h_t, v_t = w_v^T h_t \quad (8)$$

式中, h_t 为 H 中第 t 个元素, w_q, w_k, w_v 分别为将 h_t 映射到 query, key, value 的线性映射. 自注意力通过线性映射计算维度分别为 d_k, d_k, d_v 的 queries, keys, values 的值, 将它们组合为 Q, K, V 之后, 自注意力的输出如公式(9) 所示.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

在此基础之上, 为了从不同位置的不同表征子空间中挖掘信息, 进一步使用了多头注意力, 即重复执行 r 次自注意力模块生成 queries, keys, values 矩阵 $Q_i, K_i, V_i (i = 1, \dots, r)$, 计算过程如公式(10) 所示.

$$\begin{aligned} \text{MultiHead}(Q, K, V) = \\ \text{Concat}(\text{head}_1, \dots, \text{head}_r)W^O \\ \text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (10)$$

式中, W_i^Q, W_i^K, W_i^V 分别为维度为 $d_k/r, d_k/r, d_v/r$ 的并行注意力的权重矩阵, W^O 为维度为 d_o 的输出权重矩阵. CRMHSAN_AT 方法中使用了 $r = 4$ 的并行自注意力结构. 由此, 该方法可以利用序列 H 各元素之间的相对依赖关系, 逐步增强特征学习过程中情感相关信息的学习. 此外, 域不变情感特征编码器 G 最后一层是包含 128 个神经元的全连接网络层, 用于源域和目标域共享特征空间的学习.

2.3 情感分类器

情感分类器 F 是有包含两个全连接隐藏层和一层输出层的全连接网络结构, 旨在学习一个能够区分目标域语音情感的分类器. 具体说来, 在域不变情感特征编码器 G 的全连接层之后, 依次添加了有 128、64 个神经节点的全连接层. 其中最后一层隐藏层中, 使用失活率为 0.5 的 Dropout 层对神经元进行随机失活, 以此缓解模型过拟合问题. 情感分类器的损失函数是典型的多类别交叉熵, 其形式化表达如公式(11) 所示.

$$L_n^{(e)}(y^e | x; \theta_g, \theta_f^e) = -\frac{1}{n} \sum_{i=1}^n \log P(y_i^{(e)} | x_i; \theta_g, \theta_f^e) \quad (11)$$

式中, θ_g, θ_f^e 分别为共享特征编码器 G 和情感分类器 F 的参数.

2.4 域判别器

受生成模型中生成对抗网络 (Generative Adversarial Network, GAN) 的启发, 对抗训练 (Adversarial Training) 也被研究者们广泛应用于深度迁移学习任务中. Ganin 等^[8] 通过对抗训练方案来提取源域和目标域之间的域不变特征, 在特征提取器和域判别器之间添加一层梯度反转层 (Gradient Reversal Layer, GRL). 前向传播时, 梯度反转层相当于一个恒等变换函数; 反向传播时, 梯度乘以一个负系数 $-\lambda$ 用于控制反向传播. 采用伪函数对其进行形式化表述如公式(12) 所示.

$$\begin{aligned} R_\lambda(x) &= x \\ \frac{dR_\lambda}{dx} &= -\lambda I \end{aligned} \quad (12)$$

式中, I 是单位矩阵.

在 CRMHSAN_AT 中, 域判别器 D 分别由梯度反转层 GRL、包含 64 个神经元节点的全连接层和输出层构成. 其中, 全连接层之后添加一层失活率为 0.5 的随机失活层. 使用 Softmax 激活函数对输出层进行激活, 用于样本所属域标签的预测. 域判别器的损失函数如公式(13) 所示.

$$\begin{aligned} L_{n+m}^{(d)}(y^d | x; \theta_g, \theta_d) = & -\frac{1}{n} \sum_{i=1}^n \log P(y_i^d | x_i; \theta_g, \theta_d) \\ & -\frac{1}{m} \sum_{j=1}^m \log P(y_j^d | x_j; \theta_g, \theta_d) \end{aligned} \quad (13)$$

式中, θ_g 表示共享特征编码器 G 的参数, θ_d 表示域判别器 D 的参数, n, m 分别表示有情感标注的源域训练样本数和无情感标注的目标域训练样本数.

因此, CRMHSAN_AT 最终的目标函数如公式(14) 所示.

$$\begin{aligned} L(y | x; \theta_g, \theta_f, \theta_d) = & L_n^{(e)}(y^e | x; \theta_g, \theta_f) \\ & - \lambda L_{n+m}^{(d)}(y^d | x; \theta_g, \theta_d) \end{aligned} \quad (14)$$

关于 CRMHSAN_AT 的优化, 可以使用公式(15) 对网络参数进行迭代更新.

$$\begin{aligned} \theta_g &\leftarrow \theta_g - \mu \left(\frac{\partial L_n^f}{\partial \theta_g} - \lambda \frac{\partial L_{n+m}^d}{\partial \theta_g} \right) \\ \theta_f &\leftarrow \theta_f - \mu \frac{\partial L_n^f}{\partial \theta_f} \\ \theta_d &\leftarrow \theta_d - \mu \lambda \frac{\partial L_{n+m}^d}{\partial \theta_d} \end{aligned} \quad (15)$$

式中, μ 为学习率.

3 实验

为了验证 CRMHSAN_AT 方法跨语料库语音

情感识别的有效性,在两个被广泛接受和使用的英语语料库 IEMOCAP^[9] 和 MSP-IMPRO^[10] 上进行多组对照实验.

3.1 实验语料库

IEMOCAP:该语料库是在 10 个演员双向交互过程中录制的一个多模态视听语料库. 它包含了大约 12 个小时的视频动作信息和与之同步对应的音频信息(本文的研究中仅使用音频数据). 数据采集的目标是在一个可控的环境中引发出自然的情绪表达. 这一目标通过两种启发式框架实现:情感脚本和假想场景的即兴创作. 这些方法允许表演者表达由情景驱动的自发情感,而不是通过阅读来呈现典型的情感. 总计录制了 5 个双向交互过程,采用多轮标注方式地对其进行了情感标注. 每一轮都至少有两名标注者在情感属性 (Valence, Arousal 和 Dominance) 上用情感标签进行标注.

MSP-IMPRO:该数据集是一个多模态的情感语料库,记录了表演者在双向交互对话过程中的音视频. 语料库的采集依赖于一种新颖的启发式方案,在该方案中,两个演员即兴创作表演情景,其中一个演员负责说出目标句子. 对于每一个目标句子,分别创建了四种情感场景来将目标句子分别置于不同的情景中,从而引发快乐、愤怒、悲伤和中性情感. 该语料库不仅包含了目标句子,还包含了使其中一个演员说出目标句子的即兴表演中的其他句子,也包括演员休息时彼此之间的自然互动. 由 12 位演员 (6 名男性和 6 名女性) 参与录制而成,录制频率为 48 kHz,经过下采样处理之后为 44.1 kHz. 该语料库被分成 6 部分,每一部分都包含一位男演员和一位女演员. 总共由超过 9 个小时的 8 438 个情感语句组成,每个语音片段都至少有 5 名标注者对情感属性进行评估标注.

3.2 特征提取和实验设置

本本提出的方法在 Valence 维度上进行实验验证. 为了获取一致性的训练数据,对实验中使用的语料库进行了维度标签的离散化. 采用类似于文献 [11] 中的方法将维度标注转换为三种离散的情感标注,Valence 等于 3 的为中间类别,Valence 小于和大于 3 的各为一种类别,如表 2 所示. 研究发现典型的情感样本在跨语料库语音情感识别中更有价值^[12]. 因此,实验中不包含类别划分出现歧义样本,即不包括根据类别划分准则即可划分为某一类也可划分为另一类的样本.

表 2 语料库各类别样本数

语料库	低	中	高	总计
	{1, 2}	{3}	{4, 5}	
IEMOCAP	3 181	1 641	1 994	6 816
MSP-IMPRO	2 160	2 961	2 731	7 852

基准方法采用文献 [2-6] 中分别提出的 DNN、ACNN、ADD_oG、DANN、E_EDFLM 方法,实验中复现了基准模型中的语音情感识别方法. 为了与原文献中的方法保持一致,DNN、DANN、E_EDFLM 基准模型的输入特征分别采用 IS10、ComparE 2013 和 IS09 情感特征集并使用 OpenSMILE 工具包对相应的特征进行提取. 为了验证引入域对抗训练的有效性,与仅包含共享特征编码器 G 和情感分类器 F 的卷积循环多头自注意力网络 (Convolutional Recurrent Multi-Head Self-Attention Network, CRMHSAN) 进行了对比实验. CRMHSAN_AT 和 CRMHSAN 均采用与 40×751 维的 FBanks 声学特征作为网络的输入特征.

实验采用交叉验证的方式对 CRMHSAN_AT 的有效性进行验证,从目标域语料库中选取一个部分 (Session) 作为测试集对模型的性能进行测试,从源域语料库中随机选取 20% 的样本作为验证集,其余样本作为训练集组成交叉验证. 所有的实验均在 Ubuntu 16.04 LST 操作系统下完成,开发语言为 Python 3.6.2,深度学习框架为 Torch 1.3.1, GPU 为两块 GeForce RTX 2080 Ti. 公式 (14) 中的超参数 λ 设置为 0.31,优化器采用 Adam,学习率设为 0.000 1, Batch Size 设置为 64,迭代次数设置为 200,选取在验证集上损失函数最小的网络参数进行最终的性能测试.

3.3 性能评估

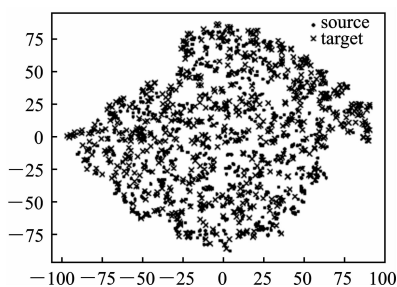
为了验证 CRMHSAN_AT 模型在跨语料库语音情感识别的有效性,与现有的跨语言或跨语料库语音情感识别方法进行了对照试验. 实验结果如表 3 所示.

表 3 跨语料库情感识别 UAR (%) 性能对比

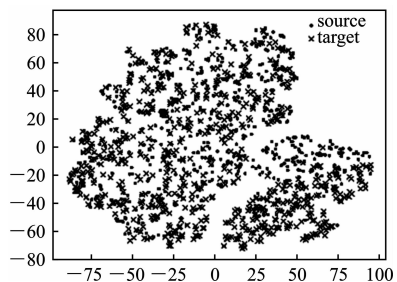
方法	源域	MSP-IMPROV	IEMOCAP
	目标域	IEMOCAP	MSP-IMPROV
IS10+DNN ^[6]		42.90	42.65
ACNN ^[5]		42.62	42.91
MADDoG ^[4]		47.40	44.40
DANN ^[3]		46.08	43.73
E_EDFLM ^[2]		46.36	44.58
CRMHSAN		45.51	43.32
CRMHSAN_AT(Ours)		47.83	45.59

由表(3)可知,CRMHSAN_AT 相比于基准模型,在带标注数据的源域数据集 MSP-IMPRO 和无标注数据的目标域数据集 IEMOCAP 上的实验表明,其测试集相对 UAR 性能提升了 0.91%~12.22%。在以 IEMOCAP 为带标注的源域语料库和以 MSP-IMPRO 为无标注的目标域语料库上的实验表明,CRMHSAN_AT 网络在测试语料库上的相对 UAR 性能比基准模型提升了 2.27%~6.90%。同时,相比于 CRMHSAN 模型,在分别以 MSP-IMPRO 和 IEMOCAP 为源域语料库上的实验结果表明,CRMHSAN_AT 方法的相对 UAR 性能分别提升了 7.46%和 5.24%。由此可知,所提出的方法能够有效地提取域不变的情感显著特征,在跨语料库语音情感识别任务上具有更好的识别性能。

为了更直观地验证本章提出的方法对域不变情感特征的学习能力。我们使用 t-SNE 工具包对 CRMHSAN_AT 方法和基线方法 CRMHSAN 的特征分布进行了二维空间映射。以 MSP-IMPRO 为源域语料库、IEMOCAP 为目标域语料库为例,图 3 展示了来自源域(“·”表示)和目标域(“×”表示)的数据分布。



(a)CRMHSAN_AT 共享特征可视化



(b)CRMHSAN 瓶颈特征可视化

图 3 源域和目标域特征分布可视化

其中,图 3(a)展示了 CRMHSAN_AT 方法中域不变情感特征编码器提取的语料库不变的情感特征分布;图 3(b)展示了 CRMHSAN 方法的瓶颈特

征分布。通过使用域对抗训练,源域和目标域样本的特征几乎是不可区分的,表明该方法能够学到一个共同的特征分布。相反,在不使用对抗训练的方法 CRMHSAN 中,特征空间中存在较大的容易辨识源域和目标域样本的区域,源域和目标域不匹配样本的存在将极大影响情感分类器的情感识别性能。实验结果表明,所提出的 CRMHSAN_AT 方法能有效地缩小不同语料库之间的差异,在目标域情感标注缺失的情况下具有更好的域不变情感特征学习能力。

4 结束语

本文提出了一种用于跨语料库的语音情感识别方法。通过不同语料库之间的对抗训练能有效地缓解不同域之间的差异对情感识别性能的影响,从而使网络学到域不变的情感显著特征。同时,使用多头注意力机制对语音序列中不同位置元素之间的相对依赖关系进行时间序列建模,增强了语音序列中情感显著部分的情感抽取能力。在两个异质化的语音情感语料库上的实验表明,在目标域情感标注缺失的情况下,所提出的方法相比于基准方法具有更好的域不变情感特征提取能力和情感识别能力。

参考文献:

- [1] LIU N, ZONG Y, ZHANG B F, et al. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5144-5148. DOI: 10.1109/ICASSP.2018.8461848.
- [2] MAO Q R, XU G P, XUE W T, et al. Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition[J]. Speech Communication, 2017(93): 1-10. DOI: 10.1016/j.specom.2017.06.006.
- [3] ABDELWAHAB M, BUSSO C. Domain adversarial for acoustic emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2423-2435. DOI: 10.1109/TASLP.2018.2867099.
- [4] GIDEON J, MCINNIS M, PROVOST E M. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (AD-DoG)[J]. IEEE Transactions on Affective Computing, 2019. DOI: 10.1109/TAFFC.2019.2916092.

[5] NEUMANN M, VU N G T. Cross-lingual and multi-lingual speech emotion recognition on English and French[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5769-5773. DOI: 10.1109/ICASSP.2018.8462162.

[6] LEE S W. The generalization effect for multilingual speech emotion recognition across heterogeneous languages[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 5881-5885. DOI: 10.1109/ICASSP.2019.8683046.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Los Angeles, USA: Curran Associates Inc., 2017: 6000-6010.

[8] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[9] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: interactive emotional dyadic motion capture database [J]. Language Resources and Evaluation, 2008, 42(4): 335-359. DOI: 10.1007/s10579-008-9076-6.

[10] BUSSO C, PARTHASARATHY S, BURMANIA A, et al. MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception[J]. IEEE Transactions on Affective Computing, 2017, 8(1): 67-80. DOI: 10.1109/TAFFC.2016.2515617.

[11] CHANG J, SCHERER S. Learning representations of emotional speech with deep convolutional generative adversarial networks [C]//Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 2746-2750. DOI: 10.1109/ICASSP.2017.7952656.

[12] SCHULLER B, ZHANG Z X, WENINGER F, et al. Selecting training data for cross-corpus speech emotion recognition: prototypicality vs. generalization [C]//Proceedings of the 2011 Afeka-AVIOS Speech Processing Conference. Tel Aviv, Israel: TUM, 2011.

作者简介:
薛艳飞 男,(1993-),硕士研究生.研究方向为语音识别.
E-mail: 2221708068@stmail. ujs. edu. cn.
张建国 男,(1964-),博士,教授.研究方向为图象处理、模式识别.

(上接第 76 页)
[11] YE S. Macronix 3V, 1Gb/2G-bit serial NAND flash memory MX35LFxGE4AB ic-stocks[R]. 2018.

作者简介:
陈亮 男,(1984-),硕士,讲师.研究方向为芯片设计、FP-

GA 技术开发工作. E-mail: Timchern@qq. com.
蔡声镇 男,(1960-),博士,教授.研究方向为嵌入式电子系统.
卢丽婉 女,(1978-),硕士,工程师.研究方向为软件技术开发.