2025 年 第 25 卷 第 11 期
2025, 25(11):04638-09

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2403487

引用格式:钟方昊, 卜凡亮, 秦昊铭. 基于单流网络的语音-人脸的跨模态学习方法[J]. 科学技术与工程, 2025, 25(11): 4638-4646.

Zhong Fanghao, Bu Fanliang, Qin Haoming. Cross modal learning method of speech face *via* single stream network[J]. Science Technology and Engineering, 2025, 25(11): 4638-4646.

基于单流网络的语音-人脸的跨模态学习方法

钟方昊, 卜凡亮*, 秦昊铭

(中国人民公安大学信息网络安全学院, 北京 100038;)

摘 要 现有的语音-人脸跨模态关联学习方法多采用双流网络结构,在降低计算复杂度、模型轻量化和高效特征融合方面还面临一些挑战,为了改善模型性能,提高跨模态学习的效率,提出一种基于单流网络的语音-人脸的跨模态学习方法。首先,将预处理的两种模态数据送入单流特征提取网络,利用基于类信息的损失函数学习提取两种模态的有效特征,接着对提取的两种模态特征向量进行基于注意力机制的特征融合,最后使用余弦相似度算法和交叉熵损失相结合的方法来学习两种模态的关联,从而完成跨模态关联学习任务。实验结果表明,本文提出的方法在语音-人脸跨模态验证、匹配和检索任务上均取得了良好的效果,在考虑网络结构轻量化和灵活性的同时保证了优秀的性能。

关键词 关联学习; 语音-人脸跨模态; 单流网络; 特征融合

中图法分类号 TP391; **文献标志码** A

Cross Modal Learning Method of Speech Face *via* Single Stream Network

ZHONG Fang-hao, BU Fan-liang*, QIN Hao-ming

(School of Information Network Security, Peoples Public Security University of China, Beijing 100038, China)

[**Abstract**] Existing methods for audio-visual cross-modal association learning often adopt a dual-stream network structure, but they still face challenges in reducing computational complexity, model light weighting, and efficient feature fusion. To improve model performance and enhance the efficiency of cross-modal learning, a single-stream network-based approach for audio-visual cross-modal learning was proposed. Firstly, preprocessed data from both modalities were fed into a single-stream feature extraction network, where a class-information-based loss function was employed to learn and extract feature vectors from both modalities. Subsequently, attention-based feature fusion was performed on the extracted feature vectors from both modalities. Finally, a combination of cosine similarity algorithm and cross-entropy loss was used to learn the association between the two modalities, thus completing the cross-modal association learning task. Experimental results demonstrate that the proposed method achieves promising performance in audio-visual cross-modal verification, matching, and retrieval tasks, ensuring excellent performance while considering the lightness and flexibility of the network structure.

[**Keywords**] association learning; voice-face cross-modal; single-stream network; feature fusion

认知科学研究人员在生物学和统计领域已经证明人的语音和面部特征具有很强的隐含关系,这一点在日常生活中也可以被主观感受到[1],显而易见的,人们可以以高于随机概率的概率将陌生的声音与不相识的面部特征进行匹配[2],根据说话人的语音信息在脑海中对说话人面部特征进行基础画像,这里的语音信息包括语义、语调、音色、音量等,代表可以从说话人语音中获得的信息[3]。给定一个人的面部图像和语音片段,利用深度学习的方法找到两种模态的对应关系,从而完成一些跨模态的任务,跨模态匹配、跨模态验证、跨模态检索等,可应用于语音生成人脸[4]视听说话人识别[5]、犯罪分析、说话人跟踪[6]和深度假视频检测[7]。

语音-人脸跨模态关联学习领域的研究,大致可归结为分类方法和度量学习方法两大类[8]。分类方法主要是将语音-人脸跨模态关联学习视为一项分类任务。SVHF(seeing voices and hearing faces)[9]是语音-人脸关联学习领域的开山之作,将跨模态1:*N*匹配问题作为分类任务来处理,创新点在于提出了一种基于卷积神经网络的模型和同时处理音

收稿日期: 2024-05-11 **修订日期**: 2024-08-01

基金项目: 中国人民公安大学安全防范工程双一流创新研究专项(2023SYL08)

第一作者: 钟方昊(2000—),男,汉族,江西赣州人,硕士研究生。研究方向:计算机视觉、多模态学习。E-mail:792999805@qq.com。

* **通信作者**: 卜凡亮(1965—),男,汉族,江苏徐州人,博士,教授,博士研究生导师。研究方向:计算机控制与信息处理。E-mail:bufanliang@sina.com。

频和视频数据新方法,解决跨模态生物识别问题,即结合音频和视频信息来识别说话者的身份。另一方面,Wen 等[10]提出了一种新模型命名为不相交映射网络(disjoint mapping network, DIMNet)。这种模型能够从诸如性别、种族、身份标识等公共变量中提取监督信号,进而学习语音和人脸的共同嵌入特征。它还能通过多个公共变量所提供的多样化标签信息,对最终的嵌入匹配效果进行评估。同时该模型运用了诸如性别、国籍等协变量,以构建语音和人脸信息之间的联系。

与分类器方案相比,度量学习方法具有更高的灵活性。因此,它已被广泛用于后续研究。随着更先进的模型结构和损耗约束的引入,现有的方法在这些任务上的结果不断改进,算法的准确性已经超越人类感官和直觉。度量学习方法的核心在于学习一个度量损失函数,目的是在公共特征空间中,使得相似的语音与人脸特征向量更加接近,不相似的则更加远离。Nagrani 等[11]采用了使用自监督学习的方式训练网络,该方法定义了一个子网络用于预测给定的人脸图像是否对应于同一视频中的语音片段,然后使用对比损失函数来训练网络,这个损失函数试图最小化同一身份的人脸和语音之间的距离,同时让不同身份的人脸和语音之间距离小于给定的阈值。Kim 等[12]利用三元组损失来学习语音和人脸的特征嵌入,并论证了这些特征向量中包含了诸如性别、年龄、种族以及面部特征的丰富信息。Horiguchi 等[13]的研究使用了一种经典的双流网络架构,用于提取语音和人脸的特征向量。运用了 N-pair 损失函数优化这两种模态特征之间的距离度量。Nawa 等[14]提出了一种新的深度训练算法,用于联合表示音频和视觉信息,作者提出一种在传统的 softmax 损失函数的基础上,增加一个基于类中心的监督信号的方法,该方法同时学习每个类别的中心和对应的图像和音频特征向量之间的距离。并且首次提出了单流网络结构,这种结构可以同时处理视觉和音频输入,不需要为每个模态单独设计网络。

语音-人脸跨模态学习领域自 2018 年提出以来,在这一领域的改进被不断提出,在相应测试指标上的性能被报告得到持续改进。但是在一致性的特征提取结构与测试流程之下,模型性能的提升是相对有限的。目前中国在这一领域的工作专注于非监督的工作,而国外的工作更致力于对于模型的性能提升和结构优化。这一领域中开源的工作是较少的,并且已开源工作所基于的数据集与测试样本往往存在较大差异,这导致与现有研究的对比面临困难。

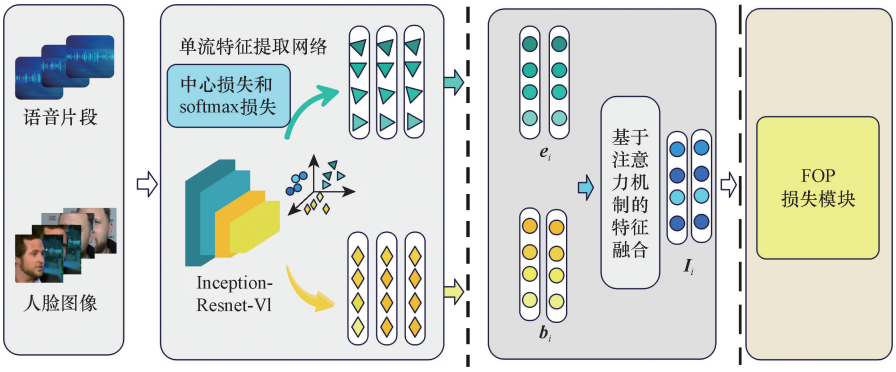
语音-人脸跨模态关联学习中,从视频数据中提取的语音和人脸信息通常存在较大噪声,另外,传统双流网络训练人脸数据和语音数据需要消耗较大的算力和资源,并且网络结构通常较为复杂,需要优化的参数较多,两种模态中蕴含的互补的说话人信息容易丢失。为了解决上述问题,本文提出了一种基于单流网络的语音-人脸的跨模态相关性学习方法。其主要工作如下:首先进行数据预处理,使用 MTCNN(multi-task convolutional neural network)[15]对人脸图片进行剪裁和对齐之后,把人脸图像统一成 256×256×3 的 RGB 图像作为单流网络的输入;同时使用信号处理技术计算语音信号的短时能量[16]进行语音数据滤波,将组合编码的音频信号根据帧速率以 16 kHz 的采样率转换为单声道[17],再将音频转化成短期幅度谱图作为单流网络的另一输入。然后使用 Inception-ResNet-V1[18]作为单流网络用于提取的音频和视觉的特征向量,通过优化跨模态中心损失函数和 softmax 损失函数,减少模态间的差异,从而提高特征提取的有效性。最后结合基于注意力机制的特征融合和正交投影损失模块(fusion and orthogonal projection, FOP)[19]来完成语音-人脸跨模态任务。

1　基于单流网络的语音-人脸的跨模态关联学习方法

本文提出的基于单流网络的语音-人脸的跨模态关联学习方法总体框架如图 1 所示,其中包括三个主要模块,首先是人脸和语音数据的单流特征提取网络;然后是基于注意力机制的特征融合模块;最后是对融合的嵌入向量进行度量学习过程的损失函数模块。

1.1　数据预处理

由于语音-人脸跨模态的数据集都来自相对应的视频数据,故采集两种模态数据的过程必然会存在一些问题使得数据的质量参差不齐,例如光照条件的变化、拍摄角度的不同,以及被采集者的姿态和表情等因素。为了提升人脸图片的质量,使用 MTCNN 对数据集进行预处理,首先确定图像中人脸的位置,通过三个级联的网络阶段(P-Net、R-Net、O-Net)来逐步精细化人脸候选框的位置,对非极大值抑制处理后的候选框进行回归和分类,从而逐步筛选出更准确的人脸区域,同时预测人脸的五个关键点位置(两个眼睛中心、鼻尖和两个嘴角),然后再对图像进行裁剪和对齐,提高人脸图像的数据质量。最后统一成 256×256×3 的人脸图片格式作为单流网络的输入。



e_i 和 b_i 分别表示语音和人脸的特征嵌入；I_i 表示特征融合后的混合向量

图 1 基于单流网络的语音-人脸跨模态关联学习方法的总体框架

Fig. 1 Overall framework of cross-modal association learning method for speech-face based on single-stream network

对于语音信号，利用计算短时能量进行语音数据滤波，首先将时变的语音信号通过使用窗函数和重叠来实现分帧，计算每一帧内所有样本的平方和作为语音数据的短时能量，然后选定阈值来区分语音段和非语音段（即静音或噪声），帧的能量低于这个阈值时，将其从语音数据中去除滤波，将保留下来的帧重新组合成一个新的语音信号以此来提高语音信号的信噪比。将组合后的音频信号根据帧速率以 16 kHz 的采样率转换为单声道，再将音频转化成短期幅度谱图作为单流网络的另一输入。

1.2 单流网络特征提取网络

传统的多网络结构通常需要为每个模态单独设计网络，这会导致计算复杂度随模态数量的增加而指数增长。设计单流网络使模型轻量化一个重要步骤，该网络可以进行端到端方式的训练，有助于网络学习到更有效的表示；另外，相对于传统的双流网络或者多网络结构，单流网络不需要人工选择成对或使用三元组信息作为监督信号，而是利用类信息来使得不同身份的表示之间的距离变远；最后，单一网络结构灵活，可以通过学习所有模态的共享表示来提高模型的泛化能力可以适用于不同模态的需求和应用场景。本文提出的单流网络是通用的，本文提出的方法使用 InceptionResNet-V1 作为单流网络，如图 2 所示，用于联合嵌入音频和视频信号，该网络使用人脸图像和频谱图进行训练。

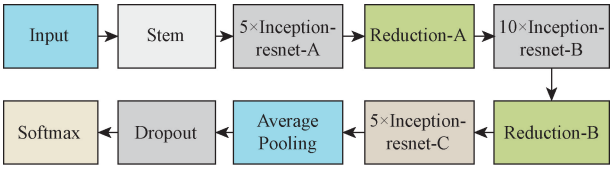


图 2 InceptionResNet-V1 结构

Fig. 2 Framework of InceptionResNet-V1

InceptionResNet-V1 主要由多个 Inception-Residual模块组成。每个 Inception-Residual 模块都包含了一个 Inception 结构和一个残差连接，其中 Stem 部分由五个 3×3 的卷积层、一个 3×3 的最大池化层和一个 1×1 的卷积层组成；InceptionResNet-A、Reduction-A 的结构如图 3 所示，其中 S 表示卷积核在输入特征图上移动的步长；InceptionResNet-B、Reduction-B 如图 4 所示，InceptionResNet-C 与 InceptionResNet-B 相似。具体来说，该模块首先通过一个 Inception 结构提取多尺度的特征，然后将这些特征与模块的输入相加，形成残差连接。这种结构既保留了 Inception 架构的多尺度特征提取能力，又利用了 ResNet 的残差连接来提高模型的训练稳定性和性能。

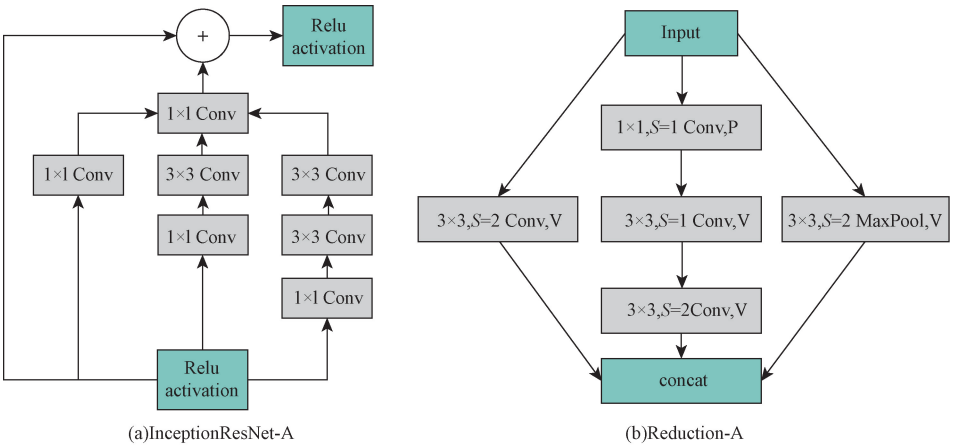
将预处理好的两种模态数据作为单流网络的输入用以提取两种模态的特征并且将音频和视觉信号映射到一个共享的潜在空间，这种网络结构能够同时处理音频和视觉信息。

假设在表示身份的类别 c 中存在与 n_i 个面部图像相关联的 n_s 个音频频谱图。每个图像和频谱图被输入网络，并且在网络的输出处获得 $n_s + n_i$ 个特征向量 $\boldsymbol{f}_c$，在训练过程中，计算 $n_s + n_i$ 个特征向量的几何中心，并使由每个特征向量到中心的距离组成的目标函数最小化，表达式为

$$d(\boldsymbol{f}_c) = \sum_{i=1}^{n_s+n_i} \left\| \boldsymbol{f}_i - \frac{1}{n_s + n_i} \sum_{j=1}^{n_s+n_i} \boldsymbol{f}_j \right\|_2^2 \tag{1}$$

式中：$\boldsymbol{f}_c$ 表示身份类别 c 中特征向量到中心距离最小的目标函数的变量，$\boldsymbol{f}^i$ 和 $\boldsymbol{f}^j$ 表示得到的特征向量。

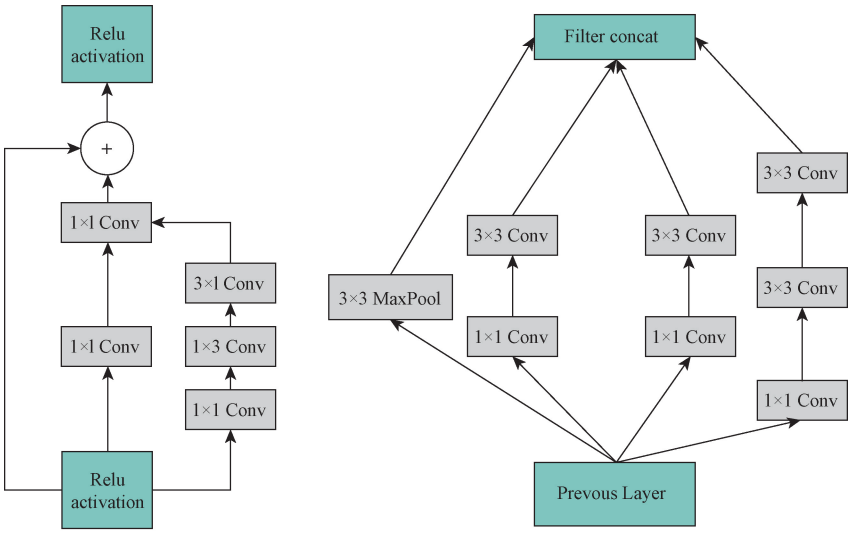
因此在训练阶段期间，单流特征提取网络将以类似的方式处理面部图像和频谱图，并且可以有效地处理图像和音频之间的差距，从而避免针对每种模态的多个网络的需要。在训练这个单流网络时，采用了跨模态中心损失函数和 softmax 损失函数相结合的方法[20]来使得网络更有效地提取特征。跨模



(a)InceptionResNet-A　　(b)Reduction-A

图 3　InceptionResNet-A 和 Reduction-A 结构图

Fig. 3　Framework of InceptionResNet-A and Reduction-A



(a)InceptionResNet-B　　(b)Reduction-B

图 4　InceptionResNet-B 和 Reduction-B 结构图

Fig. 4　Framework of InceptionResNet-B and Reduction-B

态中心损失函数旨在减少不同模态之间的差异性,使得同一身份的不同模态特征在潜在空间中更加接近。而 softmax 损失函数则用于分类任务,确保网络能够准确地识别不同的身份。这个损失函数同时学习所有类别的中心,增加了一个基于类中心的监督信号,包括小批量的人脸图像和频谱图,并最小化每个中心与相关图像和频谱图之间的距离,因此它在每个模态内以及跨模态施加邻域保持约束。如果在一个小批量 m 中有 n 个类,则该损失函数表达式为

$$L(\text{mini batch}) = -\sum_{i=1}^{m}\lg\frac{e^{\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i+b_{yi}}}{\sum_{j=1}^{n}e^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{f}_i+b_j}}+\frac{\lambda}{2}\sum_{i=1}^{m}d(\boldsymbol{f}_c) \tag{2}$$

式(2)中:$\boldsymbol{f}_i\in\mathbf{R}^d$ 为第 i 个深度特征,属于第 y_i 类;d 为特征维数;$\boldsymbol{W}_j\in\mathbf{R}^d$ 为权值的第 j 列,$\boldsymbol{W}\in\mathbf{R}^{d\times n}$ 是最后一个全连接层;$\boldsymbol{b}\in\mathbf{R}^n$ 为偏置项;参数 λ 用于平衡两个损失函数。通常情况下,将 λ 设置成 1,此时损失函数减少了类内人脸图像和频谱图之间的变化,并有效地保持了邻域结构,因此不属于同一身份的人脸图像和声谱图不会出现在同一邻域。如果 λ 被设置为 0,则只有传统的 softmax 损失函数作用于网络,可以被认为是这种联合监督的特殊情况。结合改进后的损失函数和该单流网络结构,可以得到两种模态的特征向量,并且特征向量由基于类中心的监督信号使得同一身份的特征向量相互靠近,这有利于两种模态的特征融合。

1.3　基于注意力机制的特征融合模块

多模态学习与数据融合密切相关。数据融合寻找将不同信息源组合成一个中间表示的最佳方

法,中间表示提供比单个模态更加丰富的信息。这种融合可以在不同的层次上进行,可以分为两大类:特征融合和决策融合。特征融合也称为早期融合,从不同的模态或它们的组合中寻找更好地代表解决特定问题所需的信息的特征子集。

本文中采用了基于注意力的门控多模态单元(gated multimodal units,GMU)[21]的方法,该方法是基于门控神经网络,数据融合,结合了功能和决策融合的特征融合模块,其目的是基于来自不同模态的数据的组合来找到中间表示,该模块的结构如图5所示。GMU 模型使用乘法门控来控制不同模态对隐藏单元激活的贡献,可以动态地学习不同模态对最终输出的影响程度。和注意力机制的结合使得模型可以同时学习特征表示和融合策略,而且可以从训练数据中学习到有效的门控激活模式,而不需要手动调优不同的融合权重,这种适应性使得模型在处理新的多模态数据时更加灵活。当两种模态的新样本被馈送到网络时,与两种模态相关联的门神经元接收来自所有模态的特征向量作为输入。

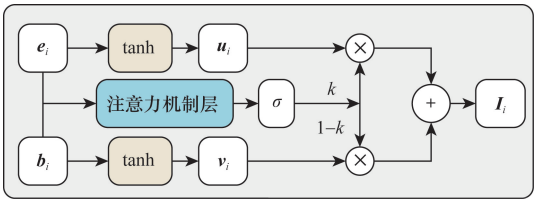


图5 基于注意力的特征融合模块结构图

Fig. 5 Structure diagram of attention-based feature fusion module

利用特征提取模块提取了面部和语音的嵌入 $\boldsymbol{b}_i$ 和 $\boldsymbol{e}_i$ 后,将人脸嵌入 $\boldsymbol{b}_i \in \mathbf{R}^F$ 投影到一个新的具有全连接层的 d 维嵌入空间 $\boldsymbol{u}_i \in \mathbf{R}^D$。类似地,将语音嵌入 $\boldsymbol{e}_i \in \mathbf{R}^V$ 投影到具有另一个全连接层的类似 d 维嵌入空间 $\boldsymbol{v}_i \in \mathbf{R}^D$,并且对 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 进行 L2 归一化,公式为

$$\boldsymbol{u}_i = \tanh(\boldsymbol{W}_u \boldsymbol{e}_i) \tag{3}$$

$$\boldsymbol{v}_i = \tanh(\boldsymbol{W}_v \boldsymbol{b}_i) \tag{4}$$

变换后的两种模态的特征向量位于一个共嵌入空间中,并且经过用于特定模态对内部表示特征进行编码 tanh 激活函数,更适合于后期的融合。

另外,对于两种模态的输入 $\boldsymbol{e}_i$ 和 $\boldsymbol{b}_i$,利用注意力机制层[22]和门神经元(在图5中由 σ 节点表示)共同作用,从两种模态的特征向量中分别计算出的单种模态特征对单元整体输出的贡献。

对于多模态融合注意力机制,从这两种模态中提取互补特征,其中一些与年龄,性别和国籍等有关,这对学习有区别的联合嵌入空间很有帮助,以此形成一个丰富的统一特征表示。多模态融合的注意力机制首先计算两种模态嵌入之间的注意力分数(亲和力),然后在用注意力分数重新校准后融合这些单个模态特征向量。计算 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 之间的注意力层的分数 α_u 和 α_v,公式为

$$\begin{aligned}\alpha_j &= \sigma(F_{\text{att}}[\boldsymbol{e}_i,\boldsymbol{b}_i]) \\ &= \sigma(\boldsymbol{W}^{\mathrm{T}}[\boldsymbol{e}_i,\boldsymbol{b}_i] + \boldsymbol{B}),\quad j \in \{\boldsymbol{u},\boldsymbol{v}\}\end{aligned} \tag{5}$$

式(5)中:σ 为 sigmoid 算子;F_{att} 为注意力层;$\boldsymbol{W}^{\mathrm{T}}$ 和 $\boldsymbol{B}$ 为注意力层学习的权重和偏置。

通常融合嵌入 $\boldsymbol{I}_i$ 的形式为

$$\boldsymbol{I}_i = \beta_u \odot \boldsymbol{u}_i + \beta_v \odot \boldsymbol{v}_i \tag{6}$$

$$\beta_j = \frac{\exp(\alpha_j)}{\sum_{i \in \{\boldsymbol{u},\boldsymbol{v}\}} \exp(\alpha_i)},\quad j \in \{\boldsymbol{u},\boldsymbol{v}\} \tag{7}$$

式中:$\odot$ 表示元素乘法;β_u 和 β_v 表示特征融合模块的输出分数。

在本文提出的方法中,满足 $\beta_v + \beta_u = 1$,因此可以用 k 和 $1-k$ 两个简单表达式来代替两种模态的注意力分数。所以对 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 进行加权之后将它们融合,计算可以得到融合嵌入 I_i 为

$$\boldsymbol{I}_i = k \odot \boldsymbol{u}_i + (1-k) \odot \boldsymbol{v}_i \tag{8}$$

1.4 损失函数

对于语音-人脸跨模态关联学习,更早的工作常采用成对或三元组损失公式来学习映射到潜在空间的嵌入,虽然取得了不错的效果,但也存在一些不足。这些方法比较依赖于距离相关的边际参数,同时程序运行时的训练复杂性较高,以及对于筛选负样本的阈值难以优化等。在本文提出的方法中,采用了计算余弦相似度和 softmax 交叉熵损失结合作为损失函数来进行度量学习,也就是 FOP 损失模块。该模块是一种轻量级、端到端的方法,该机制利用两种嵌入中的互补线索形成丰富的融合嵌入进行学习,并通过正交性约束基于它们的身份标签对它们进行聚类。

为了确保融合后的嵌入能够反映身份的语义,在嵌入空间中施加正交性约束来监督聚类,这些约束确保了来自不同身份的嵌入是正交的,而来自同一身份的嵌入则是相似的。这种方式与 softmax 交叉熵损失的角域特性更加协调,并且由于它们直接作用于批量数据,因此在训练效率上比对比损失和三元组损失的负挖掘策略更加高效。这种正交约束可以帮助学习具有辨别性的联合面部-声音嵌入,使得来自不同身份的实例在嵌入空间中更远,从而提高模型的性能。

softmax 交叉熵损失是在多分类问题中常用的损失函数,它结合了 softmax 函数和交叉熵损失的概

念。softmax 函数用于将模型的原始输出(也称为 logits)转换成概率分布,而交叉熵损失则用于衡量这个概率分布与真实概率分布之间的差异。对于本文提出的方法,融合的嵌入能够封装身份的语义,并且能够以良好的准确度预测身份标签。softmax 交叉熵损失是实现属于相同标识的实例被放置在附近,而具有不同标识标签的实例被放置在远处这一目标的最佳选择,它也允许稳定和有效的训练。具体地,该方法使用一个恒等式线性分类器,其权重表示为 $\boldsymbol{W}=[W_1,W_2,\cdots,W_C]\in\mathbf{R}^d\times C$,以此来计算对应于 $\boldsymbol{I}_i$ 的 logits。其中 d 为嵌入的维数,C 为恒等式的个数。则融合嵌入的身份分类损失计算公式为

$$L_{\mathrm{CE}}=-\lg\frac{\exp(\boldsymbol{I}_i^{\mathrm{T}}W_{y_i})}{\sum_{j=1}^{C}\exp(\boldsymbol{I}_i^{\mathrm{T}}W_j)} \tag{9}$$

由于 softmax 交叉熵损失(cross entropy loss,CEL)并不强制类别之间保持一定的间隔,这导致容易形成大小不一的类别区域,进而影响类别的可分性。有些研究尝试在欧几里得空间中引入类别间的间隔,但这种做法与交叉熵损失在角度空间实现分离的特性并不完全契合。此时本文提出的对融合嵌入实施正交约束的方法可以明确最小化同一身份内部的差异,同时最大化不同身份间的分离性。这些约束与交叉熵损失的固有角度属性相辅相成,并且由于这些约束直接在小批量数据上操作,与对比损失和三重损失所需的复杂负样本挖掘过程相比,其训练效率显著提高。其实现方式为

$$L_{\mathrm{OC}}=1-\sum_{i,j\in\boldsymbol{B},y_i=y_j}\langle\boldsymbol{I}_i,\boldsymbol{I}_j\rangle+\left|\sum_{i,j\in\boldsymbol{B},y_i\neq y_j}\langle\boldsymbol{I}_i,\boldsymbol{I}_j\rangle\right| \tag{10}$$

式(10)中:$\boldsymbol{B}$ 为小批量嵌入向量(mini batch);$\langle\boldsymbol{I}_i,\boldsymbol{I}_j\rangle$ 为余弦相似性算法,包括将嵌入投影到潜在空间进行融合和归一化处理。第一项的作用是确保同一身份内部的紧密性,即使得相同身份的融合嵌入更加聚集;而第二项则强制实现不同身份之间的分离,以保持各类别的区分度。

那么对于 FOP 损失模块,最终的损失函数为

$$L=L_{\mathrm{CE}}+\alpha L_{\mathrm{OC}} \tag{11}$$

2　实验

在 Voxceleb1[23] 数据集上验证基于单流网络的语音-人脸跨模态关联学习的有效性,具体实验情况如下。

2.1　数据集

Voxceleb1 是一个来源于 YouTube 的大规模视听人类语音视频数据集,由牛津大学在 2017 年发布。由于数据来源于真实的 YouTube 视频,因此包含了丰富的背景噪声、不同的环境条件和多样的语音风格,这使得数据集更加贴近实际应用场景。它包含了来自 1 251 位名人的 100 000 条语音数据,他们有不同的年龄、职业和口音。将数据集中的 1 225 个身份进行划分,生成了不重叠身份信息的训练集、验证集和测试集,身份个数分别为 901、100、250。

2.2　实验设置

本文所提出的方法所使用的实验平台与环境配置如表 1 所示。本文的方法是从预训练的单流网络中提取两种模态的嵌入,批量大小(mini banth)为 45,λ 设置为 1,从最后一个全连接层输出的是 512 维的特征向量。模型训练过程是在 GPU 上训练 100 轮次,分别使用了 110 281 条人脸数据和语音数据,训练的批量大小为 128,使用了具有指数衰减学习率的 Adam[24] 优化器,其学习率初始化为 10^{-5}。

表 1　实验平台与环境

Table 1　Experimental platform and environment

配置类型	参数
CPU	Intel CORE i7
GPU	NVIDIA GeForce RTX 4060 Laptop GPU
操作系统	Windows11
编程语言	Python3.9
深度学习框架	Pytorch、tensorflow
开发工具	PyCharm

2.3　实验评价指标

2.3.1　跨模态验证

跨模态验证是指给定一组语音片段和人脸图像,判断样本对是否属于同一身份,而两者之间的验证取决于相似性值的阈值,所以可以根据对真匹配的错误拒绝和对假匹配的错误接受来调整阈值,因此评价标准采用受试者工作特征曲线(receiver operating characteristic,ROC)下面积(area under curve,AUC)值和当假阳性率(false acceptance rate,FAR)等于假阴性率(false rejection rate,FRR)时的错误率(equal error rate,EER)为量化指标。

ROC 曲线是以假阳率为横轴,以真阳率为纵轴的曲线,故 AUC 值越高表示模型在验证任务上的性能越好,而 EER 值越低表示系统能够在保持较低错误率的同时保持一定的灵敏度和特异性,模型的鲁棒性和性能表现好。

另外,对于实际情况,不论是公安实践工作还是其他任务场景,跨模态验证往往是在一定的人口统计条件下,因此本文也对于相同或相似的性别(G)、年龄(A)、国籍(N)身份信息进行分组跨模态

验证任务。

2.3.2 跨模态匹配

语音-人脸跨模态匹配包括语音-人脸(voice-face,V-F)和人脸-语音(face-voice,F-V)两种情景。V-F 情景下的 1:N 匹配是指给定一段语音和 N 张人脸图像,判断这段语音和其中的哪一张人脸图像属于同一身份。F-V 情况下的 1:N 匹配是指给定一张人脸图像和 N 段语音,判断这张人脸图像和其中的哪一段语音属于同一身份。两种情况均是从 N 个样本中判断与待匹配样本属于同一身份的唯一正例,因此评价标准均采用准确性(accuracy,ACC)值为量化指标,ACC 值表示正确匹配的样本数与总样本数之比,故其值越高表示模型在匹配任务上的性能越好。

2.3.3 跨模态检索

跨模态检索任务是指给定一种模态的一个样本,从另一种模态的样本库中检索与之属于同一身份的正例,并根据样本检索结果对二者的相似度进行排序。评价标准采用平均准确率均值(meanaver age precision, mAP)为量化指标,mAP 综合考虑了准确率和排名质量,数值越高表示模型在检索任务上的性能越好。

2.4 对比实验

本文的对比实验是基于现有的具有代表性的方法与本文提出的基于单流网络的语音-人脸跨模态学习方法进行对比。

2.4.1 未分组跨模态验证

对于 voxcelceb1 数据集的未分组测试集,也就是未对测试集进行人口统计条件约束的情况下,本文提出的基于单流网络的语音-人脸跨模态学习方法和现有的具有代表性的方法进行对比实验,其结果如表 2 所示,从表 2 中可知,本文提出的方法在未分组跨模态验证任务下,降低了计算复杂度,优化了网络结构,同时保证了良好的性能,与现有效果最好的模型性能相当。

表 2 未分组跨模态验证实验结果

Table 2 Results of ungrouped cross-modal validation experiments

方法	EER/%	AUC/%
DIM Net[10]	24.9	82.5
Learnable Pins[11]	29.6	78.5
MAV-Celeb[25]	29.0	78.9
Deep Latent Space[14]	29.5	78.8
Multi-view Approach[26]	28.0	—
Adversarial-Metric Learning[27]	—	80.6
本文方法	25.6	82.5

注:“—”代表对应方法的未分组测试结果相关实验指标未经公开或者源码未开源的情况。

2.4.2 约束条件跨模态验证

对于添加约束条件的分组数据集,本文提出的基于单流网络的语音-人脸跨模态学习方法和现有的具有代表性的方法进行对比实验,其结果如表 3 所示,其中约束条件分别为性别(G)、国籍(N)和年龄(A)以及三项的组合(GNA)。实验结果表明,本文提出的方法在各个约束条件下的准确率也体现了良好的性能。

表 3 约束条件跨模态验证实验结果

Table 3 Results of cross-modal validation under constraint conditions experiments

方法	AUC/%			
	性别	国籍	年龄	GNA
DIMNet-I[10]	71.0	81.1	77.7	62.8
DIMNet-IG[10]	71.2	81.9	78.0	62.8
Learnable Pins[11]	61.1	77.2	74.9	58.8
Deep Latent Space[14]	62.4	53.1	73.5	51.4
本文方法	68.8	70.8	76.9	58.2

2.4.3 跨模态匹配

对于语音-人脸 1:N 跨模态匹配的结果如表 4 和图 6 所示,经过对实验结果的分析,可以观察到随着样本库中样本总数的增加,两种模态样本之间的匹配难度也相应增加,匹配精度持续下降。实验结果明确表明,本文所提出的方法在不同的 N 值条件下相比于现有的代表性方法仍然表现出优秀的性能。

表 4 跨模态 1:N 匹配实验结果

Table 4 Results of cross-modal 1:N matching experiments

样本数	V-F/%				
	SVHF[9]	Learnable Pins[11]	Single Stream Net[14]	Dim net[10]	本文方法
2	82	84	78	84	83
4	61	54	56	65	64
6	49	42	42	52	51
8	43	36	36	44	42
10	36	30	30	38	37

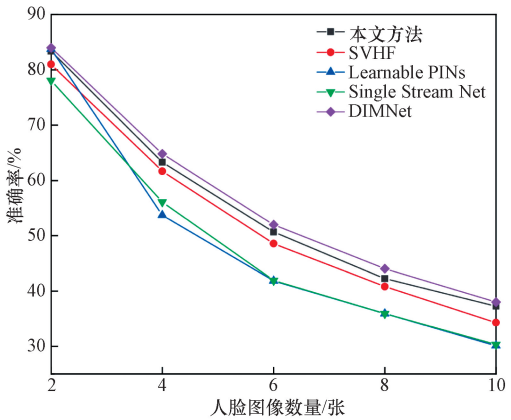


图 6 跨模态 1:N 匹配结果对比

Fig. 6 Comparison of cross-modal 1:N matching results

2.4.4　跨模态检索

跨模态检索根据给定的模态类型可以分为给定语音片段检索人脸图片(V-F)和给定人脸图片检索语音片段(F-V),由于样本库中的样本数量庞大、样本的复杂多样性使得跨模态检索任务的难度相较于其他跨模态任务显著增加。

如表5和图7所示,其中图7为跨模态检索的可视化过程。

经过对实验结果的分析,本文所提出的方法的平均mAP为7.24,在优化网络结构、降低计算复杂度的同时比现有的方法略有提升,表明了本文提出的方法对于跨模态任务具有优越性。

表5　跨模态检索实验结果

Table 5　Results of cross-modal retrieval experiment

方法	mAP/%	
	V-F	F-V
VFMR[28]	4.70	5.47
DIMNet[10]	6.22	6.65
Bi-Pcm-FST[29]	6.36	6.04
Deep Latent Space[14]	6.87	7.57
本文方法	6.92	7.55

2.5　消融实验

为了验证基于注意力机制的特征融合模块和FOP损失模块对模型性能的影响,利用Voxceleb1数据集在原实验环境和参数下进行消融实验,以未分组语音-人脸跨模态验证的实验结果为参考进行分析,实验结果如表6所示。

表6　消融实验结果

Table 6　Results of ablation experiment

消融实验条目	未分组跨模态验证实验	
	EER/%	AUC/%
单流网络+FOP损失模块	28.5	79.8
单流网络+特征融合模块+CE损失	27.5	81.6
单流网络+特征融合模块+余弦相似度损失	30.5	77.5
完整模型	25.6	82.5

从表6中可以得出结论,当删除其中任何一个模块或者采用其他损失函数时,与完整模型相比,未分组跨模态验证的准确率均有一定程度的下降。经过对实验结果的分析,单流特征提取网络使得网络更加轻量化,更具灵活性,提出的特征融合模块、FOP损失函数模块对最终结果都有一定的优化。

3　结论

目前中国越来越多的学者致力于语音-人脸跨模态学习领域的研究,包括中文语音-人脸数据集的采集、模型性能的提升、网络结构的优化以及多种应用场景的实现。国防科技大学团队[30]、人民大学团队[31]以及华侨大学[32]团队等工作在模型性能上多项评价指标都优于国外团队的工作,并且在无监督学习上开创了新的方法,有助于以公安实践为例的多种应用场景提供模型支撑,该研究领域正在成为多模态学习的热点。

本文提出了一种基于单流网络的语音-人脸关联学习方法。现有工作普遍采用双流结构,输入数据由不同结构且独立参数的两个分支子网络分别处理。这种独立分支设计增加了网络复杂性,在资源受限的计算设备中可能对模型实时性产生潜在影响。单流网络结构是一种有效的改进方案,因此本文的特征提取网络使用了单流网络结构对人脸图像和语音频谱图进行特征提取,避免了需要多个子网络来提取特征向量,简化了网络结构;同时使用了基于注意力机制的多模态门控特征融合模块,比传统的向量级联或者简单加权所得到的中间表示可以得到更加丰富的信息和两种模态之间的关联;最后使用了余弦相似度损失与交叉熵损失结合作为损失函数,余弦相似度损失函数角域特性更加协调,因此在训练效率上比对比损失和三元组损失的负挖掘策略更加高效。本文方法在跨模态验证、跨模态匹配和跨模态检索均得到了良好实验结果,表明该方法在考虑网络结构轻量化和灵活性的同时保证了优秀的性能。

图7　跨模态检索可视化结果

Fig. 7　Visualized results of cross-modal retrieval

参 考 文 献

[1] Kamachi M, Hill H, Lander K, et al. Putting the face to the voice: matching identity across modality[J]. Current Biology, 2003, 13(19): 1709-1714.

[2] Smith H M J, Dunn A K, Baguley T, et al. Matching novel face and voice identity using static and dynamic facial images[J]. Attention, Perception & Psychophysics, 2016, 78(3): 868-879.

[3] Teager H M, Teager S M. Evidence for nonlinear sound production mechanisms in the vocal tract[J]. Speech Production and Speech Modelling, 1990, 55: 241-261.

[4] 郭睿华, 宋俊鹏, 王文旭, 等. 基于视触数据融合的多模态细分类系统[J]. 科学技术与工程, 2022, 22(36): 16116-16122.
Guo Ruihua, Song Junpeng, Wang Wenxu, et al. Subdivision system based on multimodal visual and tactile data fusion[J]. Science Technology and Engineering, 2022, 22(36): 16116-16122.

[5] Tao R, Das R K, Li H. Audio-visual speaker recognition with a cross-modal discriminative network[J]. Arxiv Preprint Arxiv: 2008.03894, 2020.

[6] Liu Y, Kılıç V, Guan J, et al. Audio-visual particle flow smc-phd filtering for multi-speaker tracking[J]. IEEE Transactions on Multimedia, 2019, 22(4): 934-948.

[7] Kong C, Chen B, Yang W, et al. Appearance matters, so does audio: revealing the hidden face via cross-modality transfer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 423-436.

[8] 李俊屿, 卜凡亮, 谭林, 等. 基于多模态共享网络的自监督语音-人脸跨模态关联学习方法[J]. 科学技术与工程, 2024, 24(7): 2804-2812.
Li Junyu, Bu Fanliang, Tan Lin, et al. Self-supervised voice-face cross-modal association learning method via multi-modal shared network[J]. Science Technology and Engineering, 2024, 24(7): 2804-2812.

[9] Nagrani A, Albanie S, Zisserman A. Seeing voices and hearing faces: cross-modal biometric matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 8427-8436.

[10] Wen Y, Ismail M A, Liu W, et al. Disjoint mapping network for cross-modal matching of voices and faces[J]. Arxiv Preprint Arxiv: 1807.04836, 2018.

[11] Nagrani A, Albanie S, Zisserman A. Learnable pins: cross-modal embeddings for person identity[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 71-88.

[12] Kim C, Shin H V, Oh T H, et al. On learning associations of faces and voices[C]//Asian Conference on Computer Vision. Berlin: Springer, 2019: 276-292.

[13] Horiguchi S, Kanda N, Nagamatsu K. Face-voice matching using cross-modal embeddings[C]// Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea: ACM, 2018: 1011-1019.

[14] Nawaz S, Janjua M K, Gallo I, et al. Deep latent space learning for cross-modal mapping of audio and visual signals[C]//2019 Digital Image Computing: Techniques and Applications (DICTA). New York: IEEE, 2019: 1-7.

[15] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[16] 刘琦. 语音信号短时能量及短时幅值对比分析[J]. 网络安全技术与应用, 2011(9): 78-79.
Liu Qi. Comparative analysis of short-time energy and short-time amplitude of speech signal[J]. Journal of Network Security Technology and Applications, 2011(9): 78-79.

[17] Nagrani A, Chung J S, Zisserman A. Voxceleb: a large-scale speaker identification dataset[J]. Arxiv Preprint Arxiv: 1706.08612, 2017: 2616-2620.

[18] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[J]. Arxiv. 1602.07261, 2016.

[19] Saeed M S, Khan M H, Nawaz S, et al. Fusion and orthogonal projection for improved face-voice association[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2022: 7057-7061.

[20] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]//Computer Vision-ECCV 2016: 14th European Conference. Amsterdam: Springer International Publishing, 2016: 499-515.

[21]]Arevalo J, Solorio T, Montes-y-Gómez M, et al. Gated multimodal units for information fusion[J]. Arxiv Preprint Arxiv: 1702.01992, 2017.

[22] Chen Z, Wang S, Qian Y. Multi-modality matters: a performance leap on VoxCeleb[C]//Interspeech 2020. Slough: ISCA, 2020: 2252-2256.

[23] Nagrani A, Chung J S, Zisserman A. Voxceleb: a large-scale speaker identification dataset[J]. Arxiv Preprint, 2017: Arxiv: 1706.08612.

[24] Kingma D P. Adam: a method for stochastic optimization[J]. arXiv preprint Arxiv: 1412.6980, 2014.

[25] Nawaz S, Saeed M S, Morerio P, et al. Cross-modal speaker verification and recognition: a multilingual perspective[C]//Computer Vision and Pattern Recognition. New York: IEEE, 2021: 1682-1691.

[26] Sarı L, Singh K, Zhou J, et al. A multi-view approach to audio-visual speaker verification[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2021: 6194-6198.

[27] Zheng A, Hu M, Jiang B, et al. Adversarial-metric learning for audio-visual cross-modal matching[J]. IEEE Transactions on Multimedia, 2021, 24: 338-351.

[28] Xiong C Y, Zhang D Y, Liu Tao, et al. Voice-face cross-modal matching and retrieval: a benchmark. 2019: arXiv: 1911.09338.

[29] 朱明航, 柳欣, 于镇宁, 等. 基于双向伪标签自监督学习的跨人脸-语音匹配方法[J]. 计算机研究与发展, 2023, 60(11): 2638-2649.
Zhu Minghang, Liu Xin, Yu Zhenning, et al. Cross face-voice matching method *via* bi-pseudo label based self-supervised learning[J]. Journal of Computer Research and Development, 2023, 60(11): 2638-2649.

[30] Zhu B, Xu K, Wang C, et al. Unsupervised voice-face representation learning by cross-modal prototype contrast[J]. arXiv preprint arXiv: 2204.14057, 2022.

[31] Chen G, Zhang D, Liu T, et al. Self-lifting: a novel framework for unsupervised voice-face association learning[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval. New York: IEEE, 2022: 527-535.

[32] Wang R, Liu X, Cheung Y, et al. Learning discriminative joint embeddings for efficient face and voice association[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 1881-1884.