

基于自适应池化注意力Transformer的唇语识别方法

姚云, 胡振斌, 邓涛, 王晓
(安徽大学人工智能学院, 安徽 合肥 230031)

摘要: 唇语识别技术通过分析一系列连续的唇部图像, 建立唇部动作特征与特定语言文字之间的映射关系, 实现语义信息的识别。现有方法主要依赖循环神经网络对时序视频帧进行时序特征建模, 但存在显著的信息丢失问题, 尤其是在视频信息不完整或存在噪声干扰时, 模型往往会在区分不同时间点的唇语动作时发生混淆, 导致识别精度显著下降。针对这一问题, 提出基于自适应池化注意力Transformer的唇语识别方法 (lip reading method based on adaptive pooling attention Transformer, APAT-LR)。该方法在标准Transformer的多头自注意力 (multi-head self-attention, MHSA) 机制之前, 采用最大池化和平均池化的拼接策略, 引入自适应池化模块, 有效抑制无关信息, 增强关键特征的表达, 从而提升时序特征的建模能力。实验结果表明, APAT-LR在CMLR和GRID数据集上分别取得28.4%和1.9%的错误率, 相较于现有方法都降低了错误率, 验证了其在唇语识别任务中的有效性。

关键词: 注意力机制; Transformer; 卷积池化; 自适应

中图分类号: TP391.4

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202515

A lip reading method based on adaptive pooling attention Transformer

YAO Yun, HU Zhenxiao, DENG Tao, WANG Xiao
School of Artificial Intelligence, Anhui University, Hefei 230031, China

Abstract: Lip reading technology establishes the mapping relationship between lip movements and specific language characters by processing a series of consecutive lip images, thereby enabling semantic information recognition. Existing methods mainly use recurrent networks for spatiotemporal modeling of sequential video frames. However, they suffer from significant information loss, especially when the video information is incomplete or contains noise. In such cases, the model often struggles to distinguish between lip movements at different time points, leading to a significant decline in recognition performance. To address this issue, a lip reading method based on adaptive pooling attention transformer (APAT-LR) was proposed. This method introduced an adaptive pooling module before the multi-head self-attention (MHSA) mechanism in the standard Transformer, using a concatenation strategy of max pooling and average pooling. This module effectively suppressed irrelevant information and enhances the representation of key features. Experiments on the CMLR and GRID datasets showed that the proposed APAT-LR method could reduce the recognition error rate, thus verifying the effectiveness of the proposed method.

Key words: attention mechanism, Transformer, convolutional pooling, adaptive

0 引言

深度学习显著推动了唇语识别方法的发展。已有方法通常采用端到端模型进行训练和预测,结合计算机视觉与自然语言处理技术,系统能够自动从一系列唇部图像中提取特征,并据此生成相应的文本。卷积神经网络(convolutional neural network, CNN)^[1]和循环神经网络(recurrent neural network, RNN)^[2]分别在唇形特征提取和帧级别特征序列建模方面被广泛应用,进而捕捉唇部动作与语言之间的复杂关系。近年来,尽管深度学习方法取得了一定的进展,但仍然面临诸多挑战,特别是在长序列建模、个体差异和环境噪声带来的识别精度波动等方面。因此,在视觉编码阶段有效结合空间和时间信息,成为提升模型性能的关键之一^[3]。目前,唇语识别技术被广泛应用于多个领域,包括安全监控和身份验证^[4]、语音识别辅助^[5-6]、辅助驾驶^[7]等方面。

基于深度学习的唇语识别方法通常采用监督学习范式^[8]。该方法首先将裁剪后的唇部图像序列输入前端模型以提取视觉特征,然后将其传递至后端模型以构建时间依赖的分类模型,采用端到端方式进行训练。该方法提取用户说话时的唇形运动特征,并将其与对应的文本信息建立映射关系,能够有效实现对视频中人物言语内容的识别^[9]。

传统唇语识别编码器的视觉前端采用基于3D卷积架构或者2D卷积架构提取唇形的空间特征。Noda等^[10]首次采用由多个2D CNN堆叠而成的提取器,通过多层卷积网络自动学习图像中的空间特征,这些特征随后被用于唇语识别。与传统的手工特征提取方法相比,深度学习模型能够更好地捕捉到唇部的关键动态信息。Xu等^[11]提出了一个名为LCANet的网络,该网络专门用于编码丰富的语义特征,并在GRID语料库上进行训练,该网络的前端采用多个堆叠的3D CNN、高速公路网络提高模型捕获时空特征的能力。为降低3D对硬件资源的要求,Stafylakis等^[12]提出了3D与2D结合的特征提取器,使用灰度图像作为输入。其中3D CNN用于提取短期的唇部运动特征。2D残差神经网络(residual neural network, ResNet)有34层,包括最大池化层,用于减少特征向量的空间维度,同时捕捉丰富的唇形时空特征,提高模型的识别准确率。在对空间特征进行时序建模时,采用深度学习方法来表示不同帧之间的时序关系。Zhang等^[13]使用时间卷积网络(temporal convolutional network, TCN)进行序列建模,

克服了RNN在训练句子级唇读过程中的梯度消失问题。Ma等^[14]提出了深度可分离时间卷积网络(depthwise separable temporal convolutional network, DS-TCN),该方法使用深度可分离的时间卷积网络来减少唇读模型的参数量,并利用该模型作为教师模型进行知识蒸馏,从而优化了模型性能。

在卷积架构中,由于每个卷积核的局部感受野限制,主要提取图像局部区域的特征,如颜色、纹理或边缘等^[15]。然而,这些特征可能与唇语本身并无直接关系,反而可能干扰模型对唇形的正确识别,从而导致识别精度下降。相比之下,Transformer^[16]通过自注意力机制能够实现全局时序建模,捕捉长距离的依赖关系,但其对局部细节的捕捉能力相对较弱。考虑到视频数据中局部信息对精细特征提取至关重要,单一使用Transformer无法充分提取和理解这些重要的局部特征。为了弥补这两者的不足,研究者提出了一种基于卷积与Transformer的结合方法——自适应池化注意力Transformer。该方法首先利用卷积层提取局部空间特征,再通过自适应池化注意力模块生成动态权重向量,增强关键局部特征的表达能力。自适应池化可以根据输入的特征图大小自动调整池化的区域,有效地聚焦于最相关的局部特征,而避免不相关的局部信息对模型的干扰。

在句子级唇语识别任务中,文本解码器需面对唇形序列与文本序列长度不一致的问题。为了解决这一问题,Assael等^[17]提出了LipNet模型,采用3D卷积神经网络作为特征提取器,将双向门控循环单元(gate recurrent unit, GRU)与连续时序分类(connectionist temporal classification, CTC)结合对唇形序列进行解码。CTC是一种不需要对齐视频与文本序列的编码方法,能够处理变长序列的预测,但存在标记错位问题且缺乏文本信息的处理能力。LCANet使用级联注意力CTC解码器来处理时空特征,以捕获更长的上下文信息,但仍未解决缺乏文本信息能力问题。Chung等^[18]基于注意力机制的编解码结构提出了一个具有两个共享解码器的序列到序列模型,分别处理唇读视频特征序列和音频特征序列,生成上下文向量,并通过联合注意力模块进行文本预测,但是该方法未考虑到音频与视频对齐问题。并且,该模型采用交叉熵损失优化的是逐个字符的预测错误,训练过程中,模型关注的是最小化识别文本序列与真实文本序列之间在整个

句子层面的错误率, 这种优化目标可能与最终用于评估模型性能的标准并不完全一致, 导致目标与标准之间的差异问题。

现有解码器面临跨模态对齐模糊性与序列依赖性缺失的双重挑战^[19]: 基于CTC的模型通过blank标签缓解输入输出长度不匹配问题, 但其独立性假设导致上下文依赖断裂^[20]; 注意力机制缺少单调的对齐限制, 导致时间序列的错位问题和评价标准与优化目标不一致^[21]。因此, 本文将CTC与注意力机制结合, 提出了一种注意力增强时序连续分类网络 (attention-enhanced temporal continuity classification network, AETCCnet)。这种结合方法旨在利用CTC的自动对齐能力, 并通过注意力机制增强模型对时间步之间细粒度信息的捕捉能力。

本文的主要贡献有以下3个方面。

(1) 基于上述研究问题, 本文提出自适应池化注意力Transformer (adaptive pooling attention Transformer, APAT)。在现有的Transformer上增加自适应池化注意力模块, 实现动态调整注意力权重, 使模型更聚焦于与当前任务相关的局部细节和全局上下文, 增强Transformer对局部细节的捕捉能力, 同时保留其全局建模能力。

(2) 提出AETCCnet。在连接时序网络的基础上, 结合了注意力机制来增强时序数据的处理能力, 并且处理连续时序数据的分类任务。

(3) 基于APAT和AETCCnet, 提出基于自适应池化注意力Transformer的唇语识别方法 (lip reading method based on adaptive pooling attention Transformer, APAT-LR), 在CMLR和GRID测试集上分别达到了28.4%字符错误率 (character error rate, CER) 和1.9%的词错误率 (word error rate, WER)。相比现有模型, 其分别降低2.1%和1%的错误率。

1 基于自适应池化注意力的Transformer的唇语识别方法

人类在交流过程中常常依赖视觉信息来辅助理解语义, 尤其在嘈杂环境或听力受限的情况下, 通过观察说话者唇形的变化, 能够有效推断其表达的内容。这种能力源于大脑对唇部运动模式的直观识别和与语义的自适应关联。深度学习中的唇语识别技术正是试图模拟这一认知过程, 采用捕捉唇部时空特征的模型, 使机器具备“看唇识意”的能力^[22]。尽管现有模型能够提取唇部区域的空间特

征和时间序列信息, 但传统方法在关注关键唇部变化方面仍显不足, 难以像人类一样灵活聚焦于信息量最丰富的片段^[23]。为此, 本文提出自适应池化注意力模块 (adaptive pooling attention module, APAM), 模拟人类在观察唇形变化时自动聚焦关键帧的机制, 动态筛选出信息量最大的唇部动作进行特征提取。同时, 引入空间注意力模块, 使模型能够像人类那样专注于发音关键部位 (如双唇张合), 精准捕捉与语义紧密相关的唇部动态。两者结合显著提升了模型对唇语的感知和理解能力, 使唇语识别更接近人类的认知过程。

1.1 网络模型结构

APAT-LR网络结构如图1所示, 该模型由2个主要部分构成: 第一部分是视觉前端网络, 该模块使用3D-ResNet^[24], 用于从视频帧中提取唇部空间特征。具体来说, 3D CNN能够通过处理连续帧来捕捉时空特性, 提取出唇部在不同时刻的动态变化, ResNet 18则通过引入跳跃连接, 有效缓解了深层网络中的梯度消失问题, 保留更多关键的特征。提取唇部空间特征被进一步输入APAT编码器中, 进行时序建模, 捕捉唇部数据中的时空依赖关系。第二部分是分类网络, 基于AETCCnet架构构建, 首先通过注意力机制对视觉前端网络提取的唇部特征进行进一步融合。注意力机制的引入使模型能够自动聚焦于最关键的时间步和空间区域, 从而加强特征表达的能力, 并有效提升识别的准确度。特征融合后的数据将经过全连接层映射到分类空间, 随后通过Softmax函数进行标准化处理, 确保特征在分类过程中具有一致性。最终, 这些经过归一化后的特征被输入CTC网络, 完成模型的训练和解码过程, 并实现从唇部动作到语言内容的精确预测。在模型预测阶段, 使用束搜索进行解码, 以提高序列预测的准确性和鲁棒性。

1.2 APAT编码器

APAT编码器结构如图2所示。其中, APAM将视觉前端提取的帧级别特征作为输入序列 $X \in \mathbb{R}^{B \times T \times F}$, B 表示批次, T 表示视频帧数, F 表示特征维度。

APAT编码器由多个APAT层堆叠而成, 每个APAT层包含了APAM、多头自注意力 (multi-head self-attention, MHSA) 机制^[25]、前馈网络 (feed-forward network, FFN) 和层归一化 (layer normalization, LN)。这些模块共同构成了一个高效的特

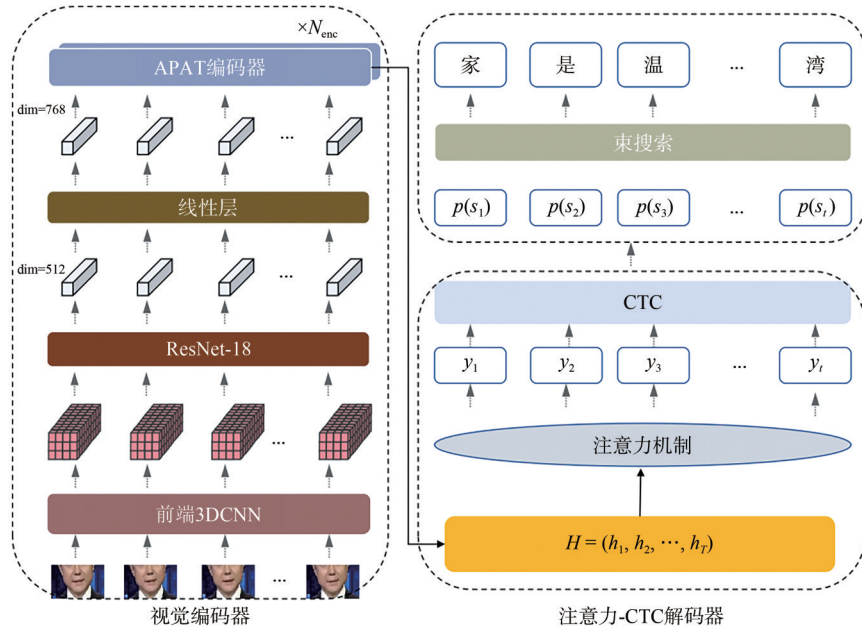


图1 APAT-LR网络结构

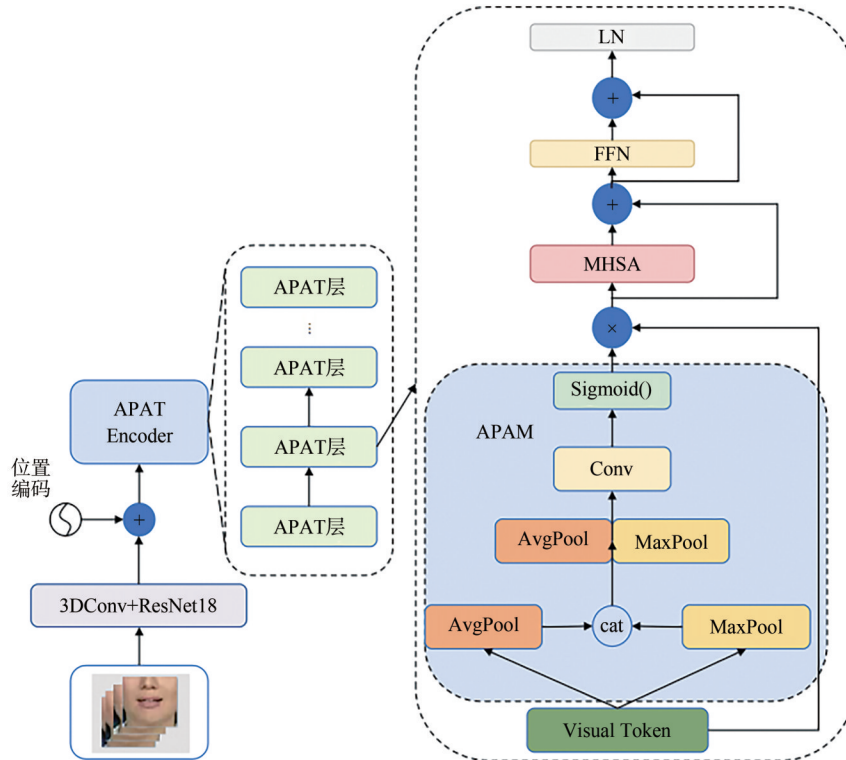


图2 APAT编码器结构

征提取和建模框架，能够精确捕捉输入数据中的局部细节及全局上下文信息。

APAM 首先对输入特征 X 使用基于特征维度的最大池化和平均池化操作，得到全局通道描述符：

$$F_{\max} = \text{MaxPool}(X) \quad (1)$$

$$F_{\text{avg}} = \text{AvgPool}(X) \quad (2)$$

接着， $F_{\max}$ 和 F_{avg} 在特征维度上拼接，使通道数变为2，使用卷积核大小为 1×1 的2D CNN 进行降维，将通道数压缩为1。通过 Sigmoid 函数的激活，生成时空维度上的注意力权重 A_j ：

$$A_J = \sigma \left(\text{Conv}_{1 \times 1} \left(\left[F_{\max}; F_{\text{avg}} \right] \right) \right) \quad (3)$$

该权重与原始特征 X 相乘, 使时空级别特征进行选择强化:

$$X' = A_J \odot X \quad (4)$$

增强后的特征 X' 输入多头自注意力模块, 以捕捉全局依赖关系, 首先将 X' 通过线性变换得到 query (Q)、key (K) 和 value (V)。

自注意力权重的计算过程如下:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (5)$$

多头注意力将输入线性投影到 h 个头, 并对每个头独立计算注意力, 最终拼接各个头输出并线性变换:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W^O \quad (6)$$

其中, $W^O \in \mathbb{R}^{d_v \times d_v}$ 为输出投影矩阵。

多头自注意力模块的输出经过残差连接和LN, 以稳定训练并增强特征表达能力:

$$X'' = \text{LayerNorm}(X' + \text{MultiHead}(Q, K, V)) \quad (7)$$

每个 APAT Layer 包含一个 FFN, 由两层全连接层和 GELU 激活函数构成:

$$\text{FFN}(x) = \text{FC}(\text{GELU}(\text{FC}(x))) \quad (8)$$

输出经过残差连接和LN:

$$X''' = \text{LayerNorm}(X'' + \text{FFN}(X'')) \quad (9)$$

经过 N 层 APAT Layer 堆叠后, 充分建模视觉的时空特征, 最终输出特征的计算过程如下:

$$Z_N = \text{APATLayer}_N(\text{APATLayer}_2(\text{APATLayer}_1(X))) \quad (10)$$

经过归一化处理的特征用于下游任务, APAT 编码器利用 APAM 提取关键帧信息, 并结合 MHSA 机制捕捉全局依赖, 增强模型的时空特征。引入 MHSA 机制, 以并行方式捕捉序列中不同时间步之间的全局依赖关系。同时, 在编码过程中融合了位置编码信息与跨帧上下文线索, 使得提取的特征更具时序结构性和语义一致性。编码后的高层次特征表为后续的分类器提供了稳定、判别性强的输入基础。

1.3 注意力增强时序连续分类网络

CTC 通常与具有时序建模能力的神经网络结合使用, 其主要目标是优化 GRU^[26]等网络的内部

参数, 以最大化输出标注序列的对数概率。CTC 网络基于编码器输出的隐藏状态序列, 对每个时间步的隐藏状态进行处理, 通过 Softmax 层计算每个时间步对应标签的概率分布。

编码器输出的隐藏状态序列 $H = (h_1, h_2, \dots, h_T)$, 其对应的 Softmax 层输出:

$$y_t = \text{Softmax}(Wh_t + b) \quad (11)$$

其中, W 和 b 为可学习参数, y_t 在时刻 t 上各个标签的概率分布。

给定标签序列 l , 其对应的概率分布为所有可能路径的概率和:

$$P(l|x) = \sum_{\pi \in B^{-1}(l)} P(\pi|x) \quad (12)$$

其中, π 表示所有可能的路径, $B^{-1}(l)$ 是去除空白符和重复符号后得到的标签序列的路径集合。

CTC 使用负对数似然作为损失函数:

$$L_{\text{CTC}} = -\log P(l|x) \quad (13)$$

为进一步提升模型性能, 本文提出了融合注意力机制与 CTC 解码的 AETCCnet 网络。该方法在解码阶段利用注意力机制更好地对齐编码器输出特征与目标序列。融合注意力机制和连接时序分类解码网络的结构如图3所示。

编码器使用注意力机制进行相似度计算, 解码器在每个时间步长计算当前隐藏状态 s_t 与编码器输出 h_i 。

$$e_{it} = v^T \tanh(W_h h_i + W_d s_t + b_a) \quad (14)$$

其中, W_h 、 W_d 和 v 表示可学习参数。通过 Softmax 函数计算注意力分布:

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^T \exp(e_{t,k})} \quad (15)$$

利用注意力权重对编码器输出进行加权求和, 得到上下文向量:

$$C_t = \sum_{i=1}^T a_{t,i} h_i \quad (16)$$

上下文向量与当前隐藏状态结合, 通过全连接层映射到标签空间, 并使用 Softmax 函数计算最终输出的概率分布:

$$P(y_t | y_{<t}, x) = \text{Softmax}(W_o[s_t; c_t] + b_o) \quad (17)$$

在预测阶段, 模型选择每个时间步上概率最大的标签:

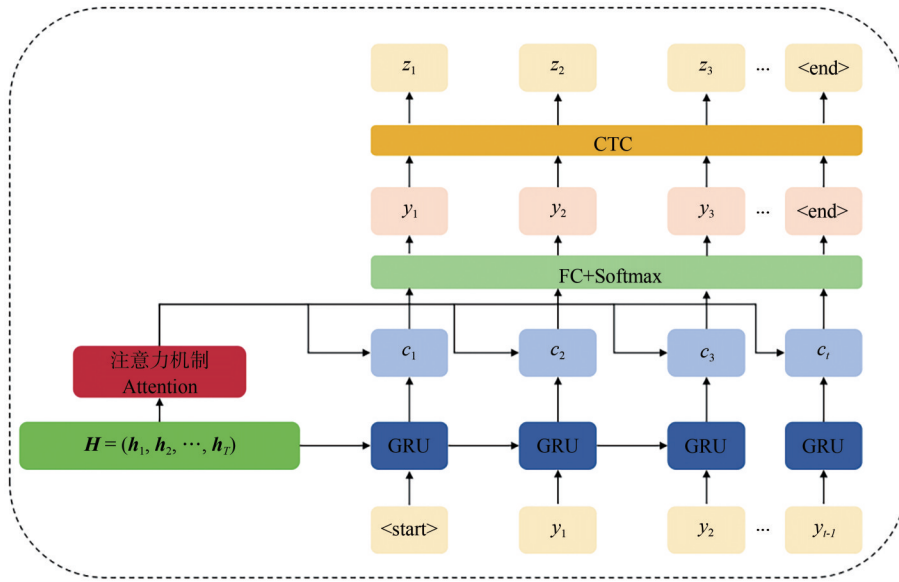


图3 融合注意力机制和连接时序分类解码网络的结构

$$y_t = \arg \max_k P(k | x_t) \quad (18)$$

最后使用CTC解码规则（合并连续重复字符并移除空白符号）得到最终的输出序列。

2 实验过程及其分析

2.1 数据集

本文的研究工作在CMLR^[27]和GRID^[28]两个数据集上进行了实验。CMLR数据集由浙江大学视觉智能与模式分析组构建，旨在推动唇语识别领域的深入研究。该数据集收集了2009年6月至2018年6月中国国家新闻节目《新闻联播》中11位发言人的102 072个口语句子。每个句子最长可达29个汉字，且不包含英文字母、阿拉伯数字或罕见标点符号。CMLR数据集还提供了每个词的时间对齐边界信息（以秒为单位），以便于精细的时序分析。GRID数据集是唇语识别领域广泛使用的标准化英文句子级数据集，共包含32 823个视频片段，涉及34位说话人。每段视频固定为75帧，句子均由6个单词组成，这些单词来自预定义的有限词汇表，为唇语识别任务提供了高度结构化的数据基础。

唇语识别性能通常采用CER和WER作为评价指标。CER和WER分别衡量汉字字符和单词在预测结果与真实标签之间的编辑距离。CER用于衡量模型在字符级别上的识别准确性，数值越小表示识别效果越好，其数学表达式如下：

$$CER = \frac{D+S+I}{N} \quad (19)$$

其中， D 、 S 和 I 分别表示将预测序列转换为真实标签过程中所需的删除、替换和插入的汉字个数， N 为真实标签中汉字的总数。

2.2 数据预处理

本文首先将视频以25帧/秒的速率转换为连续的图片帧。采用RetinaFace Predictor方法^[29]进行人脸检测，精确定位人脸的大致区域。接着，使用FAN Predictor方法^[30]进行面部关键点的细粒度检测，并追踪每帧中人脸的68个关键标志点，这些点主要分布在眼睛、嘴唇和鼻子等面部区域。基于这些标志点，从每帧图像中裁剪出尺寸为64像素×128像素的唇部区域。裁剪完成后，将唇部图像转换为灰度图，以简化数据并减少计算复杂度。最后，对灰度图进行归一化处理，利用训练集的均值和方差对图像进行标准化，以确保数据的一致性与可比性。数据预处理的完整流程如图4所示。

2.3 实验环境与配置

本文实验使用的硬件配置包括12代Intel(R) Core(TM) i5-12600KF处理器、32 GB运行内存及NVIDIA GeForce RTX 3060 Ti显卡。软件环境基于64位Ubuntu 18.04操作系统，搭建于Pytorch深度学习框架下。本文实验采用的硬件和软件环境为深度学习模型的训练和测试提供了强大的支持，确保了实验的顺利进行和结果的可靠性。

2.4 数据增强

本文采用随机裁剪作为数据增强方法^[31]。该方法涉及在原始图像上随机选取一个坐标点，以

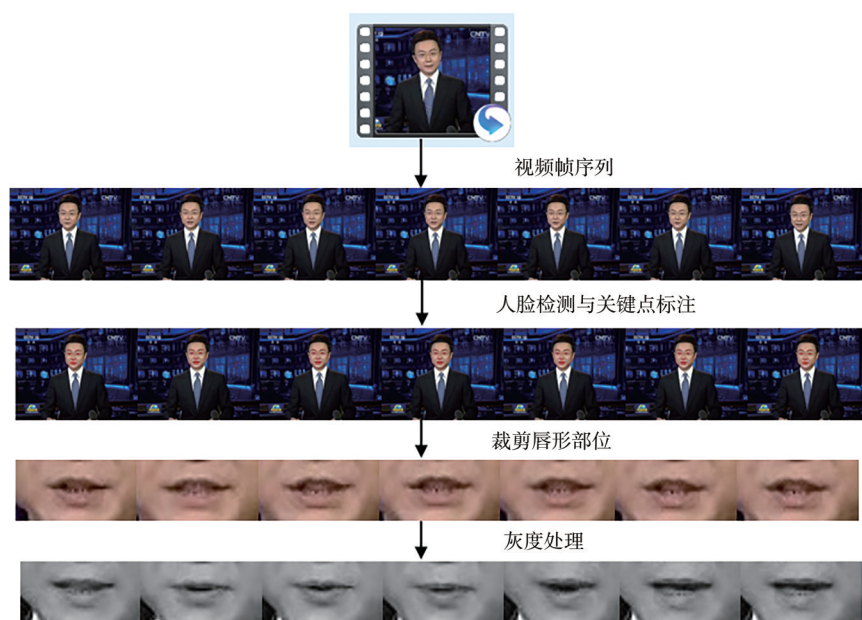


图4 数据预处理流程

此点作为裁剪区域的起始左上角，进而从这个坐标点截取出一个 $88\text{ px} \times 88\text{ px}$ 的图像片段，通过这种方式，加强模型对图像中目标对象的定位和识别能力。

2.5 训练设置

模型训练过程中采用 Adam 优化器^[32]，初始学习率设置为 0.0003，并引入了学习率调度策略以优化训练效果。在前 25 000 个训练步骤内，学习率保持在峰值 0.0003，以确保模型充分学习特征；随后，学习率按照训练步数平方根的倒数逐步衰减，防止过拟合并提升模型的泛化能力。训练采用小批量梯度下降，共训练 100 个 epochs，批量大小设置为 5，以确保模型在数据上得到充分优化。为了加速训练过程并提高模型的稳定性，在模型的各个层之间应用了批量归一化技术。为了减少模型对特定训练样本的过度依赖，在网络层中添

加 Dropout 层，该层会随机丢弃一部分神经元的输出。

2.6 与现有方法的对比

本文对比了不同方法在 CMLR 数据集和 GRID 数据集上的 CER 和 WER，结果见表 1。根据所使用的时序网络的不同大致可以分为 LSTM、GRU、TCN。本文提出的 APAT-LR 方法在 CMLR 和 GRID 的测试集上的 CER 和 WER 低于其他方法，通过使用 APAT 编码器和 AETCCnet 的组合在 CMLR 测试集和 GRID 测试集上获得了最佳的识别效果。

2.7 消融实验

为了验证本文使用的数据增强方法以及方法模块的有效性，对数据增强方法、APAT 编码器以及 AETCCnet 进行了消融实验，结果见表 2。其中，时序特征提取使用 APAT 编码器，分类网络采用 AETCCnet。通过消融实验结果可以观察到，不同

表1 本方法和其他方法在两个数据集的指标对比

模型	视觉前端	后端分类网络	CMLR 上的 CER	GRID 上的 WER
LipCH-Net ^[17]	LipCH-Net	RNN+CTC	42.1%	4.3%
WAS ^[18]	VGG	GRU(attention)	39.93%	3.1%
WLAS ^[33]	3D Conv+ResNet 18	LSTM(attention)	39.4%	2.7%
CSSMCM ^[27]	VGG	GRU(attention)	38.2%	2.3%
CHSLR-VP ^[34]	CNN+ResNet 18	GRU+MLP	35.2%	2.4%
LIBS ^[35]	VGG-M	RNN(attention)+MLP	31.27%	2.31%
3DT-CHLipNet ^[36]	3D Conv+ResNet 18	TCN+CTC	30.31%	2.2%
APAT-LR	3D Conv+ResNet 18+APAT	AETCCnet	28.4%	1.9%

模块对识别准确率的提升都有帮助,同时使用时可帮助模型获得更好的识别效果。

表2 消融实验结果

数据增强 模块	APAT 模块	AETCCnet 模块	CMLR 上的 CER	GRID 上的 WER
√	×	×	40.5%	3.50%
×	√	×	38.9%	3.23%
×	×	√	35.1%	3.11%
√	√	×	33.3%	2.65%
√	×	√	32.2%	2.61%
×	√	√	31.3%	2.55%
√	√	√	28.4%	1.90%

2.8 可视化分析

为了验证本文提出的 APAT-LR 在处理唇形序列时的有效性,实验分析了其在不同时间步长下的显著性图^[31]。唇形显著性图如图5所示。本文选取某一个视频的若干帧进行显著性可视化,通过对比发现,在加准模型增加 APAT 模块时,模型在单一帧中更能重点关注唇形区域。对于唇形序列,在唇形序列处理中,该模型能够识别并聚焦于关键帧,从而有效提高了识别性能。

3 结束语

本文针对唇语识别中的时序建模问题,提出了一种名为 APAT-LR 的唇语识别方法。具体而言,该方法通过 3D CNN+ResNet 视觉前端直接从原始像素中提取特征,然后将这些特征输入 APAT 编码器进行时序建模。在 APAT 编码器的输出上,使用 AETCCnet 解码器进行汉字预测。在模型训练过程中,本文结合 CTC 与注意力机制,采用混合训练

策略来提升模型性能。在 CMLR 和 GRID 数据集上的实验结果表明,APAT-LR 方法在 CER 这一评估指标上表现优于现有方法,证明了其有效性和优越性。

该技术具有广泛的应用潜力,可适用于多个领域和场景:在无障碍交互场景中,可赋能听障人士通过智能设备实现实时唇语转文字交互,尤其在嘈杂或静音环境中其具有不可替代性;在公共安全领域,其能通过监控视频还原关键对话内容,为司法取证和反恐情报分析提供技术支撑。未来研究可围绕低资源方言适配、边缘计算部署优化及多模态情感分析等方向深化,同时需建立数据隐私保护框架以规避技术滥用风险。本研究的技术成果不仅推动了中文唇语识别的理论发展,更为构建“无声智能交互生态”提供了关键技术支撑,具有显著的学术价值与社会效益。

现有的 CMLR 和 GRID 数据集在背景环境方面较为单一,通常是在受控环境下录制的视频数据,这使得模型在复杂真实场景中的适应性和鲁棒性受到限制。由于这些数据集缺乏对不同光照条件、摄像机角度变化、背景干扰因素(如噪声、其他人物或动态背景)的充分覆盖,模型在面对复杂现实环境时的泛化能力仍存在较大挑战。因此,未来的研究工作将重点探索在多样化背景下的唇语识别性能,包括引入更多样化的拍摄场景、增强数据集的多样性,甚至结合数据增强或生成对抗网络等方式来提升模型在不同背景下的适应能力。此外,还将考虑跨数据集的迁移学习,以提高模型在未见过的场景中的鲁棒性,从而推动唇语识别技术在真实应用场景中的落地和发展。

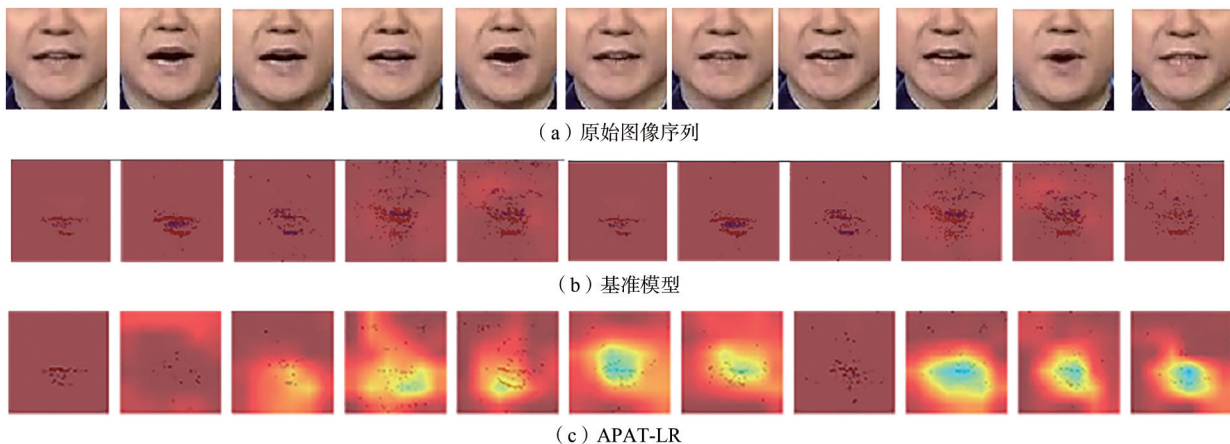


图5 唇形显著性图

参考文献:

- [1] 曾淦雄, 柯道. 基于3D卷积的图像序列特征提取与自注意力的车牌识别方法[J]. 智能科学与技术学报, 2021, 3(3): 268-279.
ZENG G X, KE X. 3D convolution-based image sequence feature extraction and self-attention for license plate recognition method[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(3): 268-279.
- [2] 蔡莹皓, 杨华, 安璇, 等. 神经符号学及其应用研究[J]. 智能科学与技术学报, 2022, 4(4): 560-570.
CAI Y H, YANG H, AN X, et al. Study on NeuroSymbolic learning and its applications[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(4): 560-570.
- [3] SUMBY W H, POLLACK I. Visual contribution to speech intelligibility in noise[J]. The Journal of the Acoustical Society of America, 1954, 26(2): 212-215.
- [4] EXARCHOS T, DIMITRAKOPOULOS G N, VRAHATIS A G, et al. Lip-reading advancements: a 3D convolutional neural network/long short-term memory fusion for precise word recognition[J]. BioMedInformatics, 2024, 4(1): 410-422.
- [5] WENG X S, KITANI K. Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading[J]. arXiv preprint, 2019, arXiv: 1905.02540.
- [6] TATULLI E, HUEBER T. Feature extraction using multimodal convolutional neural networks for visual speech recognition[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2017: 2971-2975.
- [7] MALEK M, BEN M M A, AICHA B. Lip reading of words with lip segmentation and deep learning[J]. Multimedia Tools and Applications, 2022, 82(1): 551-571.
- [8] 谢胤岑, 薛峰, 曹明伟. 基于多重视觉注意力的唇语识别[J]. 模式识别与人工智能, 2024, 37(1): 73-84.
XIE Y C, XUE F, CAO M W. Lip recognition based on multiple visual attention[J]. Pattern Recognition and Artificial Intelligence, 2024, 37(1): 73-84.
- [9] 吴威龙, 李润恺, 许霜桦, 等. 基于自适应序列帧长度的端到端式唇语识别算法[J]. 生命科学仪器, 2023, 21(4): 35-39.
WU W L, LI R K, XU S Y, et al. An end-to-end lip-reading recognition algorithm based on the adaptive length of frame sequence[J]. Life Science Instruments, 2023, 21(4): 35-39.
- [10] NODA K, YAMAGUCHI Y, NAKADAI K, et al. Lipreading using convolutional neural network[C]//Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2014). Singapore: ISCA, 2014: 1149-1153.
- [11] XU K, LI D W, CASSIMATIS N, et al. LCANet: end-to-end lipreading with cascaded attention-CTC[C]//Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Piscataway: IEEE Press, 2018: 548-555.
- [12] STAFYLAKIS T, TZIMIROPOULOS G. Combining residual networks with LSTMs for lipreading[J]. arXiv preprint, 2017, arXiv:1703.04105.
- [13] ZHANG T, HE L, LI X D, et al. Efficient end-to-end sentence-level lipreading with temporal convolutional networks[J]. Applied Sciences, 2021, 11(15): 6975.
- [14] MA P C, MARTINEZ B, PETRIDIS S, et al. Towards practical lipreading with distilled and efficient models[C]//Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 7608-7612.
- [15] PENG Z Q, LUO Y H, SHI Y, et al. SelfTalk: a self-supervised communicative training diagram to comprehend 3D talking faces[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 5292-5301.
- [16] 丛晓峰, 桂杰, 章军. 基于视觉Transformer的多损失融合水下图像增强网络[J]. 智能科学与技术学报, 2022, 4(4): 522-532.
CONG X F, GUI J, ZHANG J. Underwater image enhancement network based on visual Transformer with multiple loss functions fusion[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(4): 522-532.
- [17] ASSAEL Y M, SHILLINGFORD B, WHITESON S, et al. LipNet: end-to-end sentence-level lipreading[J]. arXiv preprint, 2016, arXiv: 1611.01599.
- [18] CHUNG J S, SENIOR A, VINYALS O, et al. Lip reading sentences in the wild[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 3444-3453.
- [19] XUE F, YANG T, LIU K, et al. LCSNet: end-to-end lipreading with channel-aware feature selection[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(1s): 1-21.
- [20] 洪依, 孙成立, 冷严. 基于超轻量通道注意力的端到端语音增强方法[J]. 智能科学与技术学报, 2021(3): 351-358.
HONG Y, SUN C L, LENG Y. End-to-end speech enhancement based on ultra-lightweight channel attention[J]. Chinese Journal of Intelligent Science and Technology, 2021(3): 351-358.
- [21] BURCHI M, TIMOFTE R. Audio-visual efficient conformer for robust speech recognition[C]//Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2023: 2257-2266.
- [22] LU J C, SISMAN B, LIU R, et al. VisualTts: TTs with accurate lip-speech synchronization for automatic voice over[C]// Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 8032-8036.
- [23] SHI B W, HSU W N, LAKHOTIA K, et al. Learning audio-visual speech representation by masked multimodal cluster prediction[J]. arXiv preprint, 2022, arXiv:2201.02184.
- [24] WANG H J, CUI B Y, YUAN Q B, et al. Mini-3DCvT: a lightweight lip-reading method based on 3D convolution visual transformer[J]. The Visual Computer, 2025, 41(3): 1957-1969.
- [25] LI Z Y, LOHRENTZ T, DUNKELBERG M, et al. Transformer-based lip-reading with regularized dropout and relaxed attention[C]//Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE Press, 2023: 723-730.
- [26] HE L, DING B Y, WANG H, et al. An optimal 3D convolutional neural network based lipreading method[J]. IET Image Processing, 2022, 16(1): 113-122.
- [27] ZHAO Y, XU R, SONG M L. A cascade sequence-to-sequence model for Chinese mandarin lip reading[C]// Proceedings of the 1st ACM International Conference on Multimedia in Asia. New York: ACM, 2019: 1-6.
- [28] COOKE M, BARKER J, CUNNINGHAM S, et al. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2006, 120(5 Pt 1): 2421-

2424.

- [29] YANG L L, XU X S, ZHAO J Z, et al. Fusion of RetinaFace and improved FaceNet for individual cow identification in natural scenes[J]. Information Processing in Agriculture, 2024, 11(4): 512-523.
- [30] BULAT A, TZIMIROPOULOS G. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230, 000 3D facial landmarks)[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 1021-1030.
- [31] ZHANG X Y, ZHANG C W, SUI J P, et al. Boosting lip reading with a multi-view fusion network[C]//Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2022: 1-6.
- [32] PRAJWAL K R, AFOURAS T, ZISSERMAN A. Sub-word level lip reading with visual attention[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 5152-5162.
- [33] MA P C, PETRIDIS S, PANTIC M. Visual speech recognition for multiple languages in the wild[J]. Nature Machine Intelligence, 2022, 4(11): 930-939.
- [34] 何珊, 袁家斌, 陆要要. 基于中文发音视觉特点的唇语识别方法研究[J]. 计算机工程与应用, 2022, 58(4): 157-162.
- HE S, YUAN J B, LU Y Y. Research on lip reading based on visual characteristics of Chinese pronunciation[J]. Computer Engineering and Applications, 2022, 58(4): 157-162.
- [35] ZHAO Y, XU R, WANG X C, et al. Hearing lips: improving lip reading by distilling speech recognizers[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 6917-6924.
- [36] 刘培培, 贾静平. 基于时域卷积网络的中文句子级唇语识别算法[J]. 计算机应用研究, 2023, 40(9): 2596-2602.
- LIU P P, JIA J P. Chinese sentence-level lip recognition algorithm based on time domain convolutional network[J]. Application Research of Computers, 2023, 40(9): 2596-2602.

[作者简介]



姚云 (1996-), 男, 安徽大学人工智能学院硕士生, 主要研究方向为唇语识别、生成式人工智能、计算机视觉。



胡振琥 (2000-), 男, 安徽大学人工智能学院硕士生, 主要研究方向为异常行为识别、计算机视觉、深度学习。



邓涛 (2003-), 男, 安徽大学人工智能学院本科生, 主要研究方向为计算机视觉、自动驾驶、智能自主系统。



王晓 (1988-), 女, 安徽大学人工智能学院教授, 主要研究方向为社会计算、群体行为建模、无人自主系统及其平行测试。