

Equation Chapter 1 Section 1 基于自适应池化注意力的 Transformer 的唇语识别方法

姚云, 胡振, 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 的多头自注意力机制 (Multi-Head Self-Attention, MHSA) 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 (Adaptive Pooling Module), 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。该实验结果表明, APAT-LR 在 CMLR 和 GRID 数据集上分别取得 28.4% 和 1.9% 的字符错误率, 相较于现有方法分别降低了错误率, 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:

文献标志码:

The 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) is proposed. This method introduces 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 helps suppress irrelevant information and enhances the representation of key features. Experiments on the CMLR and GRID datasets show that the proposed APAT-LR method can reduce the recognition error rate, thus verifying the effectiveness of the proposed method.

Key words: Attention Mechanism, Transformer, Convolutional Pooling, Adaptive

1 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(CNN)和循环神经网络(RNN)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,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,在视觉编码阶段如何有效结合空间和时间信息,成为提升模型性能的关键问题之一[1]。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[2]、语音识别辅助[3-4]、辅助驾驶[5]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[6]。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。通过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,该方法能够有效实现对视频中人物言语内容的识别。

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等人[7]首次采用由多个2D卷积神经网络(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等人[8]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卷积网络和高速公路网络提高模型特征空间特征的能力。为降低3D对于硬件资源的要求,Stafylakis和Tzmiropoulos等人[9]提出了3D卷积与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 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[10]人使用时间卷积网络(TCN)进行序列建模,克服RNN在训练句子级唇读过程中的存在的梯度消失问题。Ma等人[11]提出了深度可分离时间卷积网络(DS-TCN),该方法通过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。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[12]通过自注意力机制(Self-Attention)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于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,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,并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,文本解码器面对唇形序列与文本序列的长度通常不一致的问题。为了解决这一问题,Assael等人[13]提出了LipNet模型,采用3D卷积神经网络(3D CNN)作为特征提取器,将双向门控循环单元(GRU)与CTC(Connectionist Temporal Classification)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,同时存在标记错位问题和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等人[14]基于注意力机制的编解码结构提出的通过一个具有两个共享解码器的序列到序列(Sequence-to-Sequence)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。同时采用交叉熵损失优化的是逐个字符的预测错误,

而 WER 优化的是整个句子层面的错误, 在训练过程中存在优化目标与评估标准不同的问题。

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[15]: 基于 CTC 的模型通过空白标签 (blank) 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; 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的问题。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, 同时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:

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。

提出注意力增强时序连续分类网络 (Attention-Enhanced Temporal Continuity Classification Network, 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。

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% 和 1.9% 的字符错误率, 相比现有模型, 分别降低 2.1% 和 1% 的字符错误率。

2 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, 通过捕捉唇部时空特征的模型, 使机器具备“看唇识意”的能力。

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。为此, 本文提出自适应池化注意力模块 (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, 使模型能够像人类专注于发音关键部位 (如双唇闭合、张合等), 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。

2.1 网络模型结构

该模型由两个主要部分构成: 第一部分是视觉前端网络, 该模块使用 3D-ResNet, 用于从视频帧中提取唇部空间特征。具体来说, 3D 卷积神经网络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, 残差网络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到 APAT (自适应池化注意力 Transformer) 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, 它基于 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得模型能够自动聚焦于最为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将被输入到 CTC (连接时序分类) 网络中, 完成模型的训练和解码过程, 最终实现从唇部动作到语言内容的精确预测。

2.2 APAT 编码器

APAT 编码器的结构如图 2 所示, 其中, APAM 代表自适应池化注意力模块, 该模块以视觉前端提取的帧级别特征作为输入序列 $X \in \mathbb{R}^{B \times T \times F}$, 其中 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 Layer 堆叠而成, 每个 APAT Layer 包含了自适应池化注意力模块 (APAM)、多头自注意力机制 (Multi-Head Self Attention)、前馈网络 (FeedForward Network) 和层归一化 (LayerNorm, LN)。这些模块共同构成了

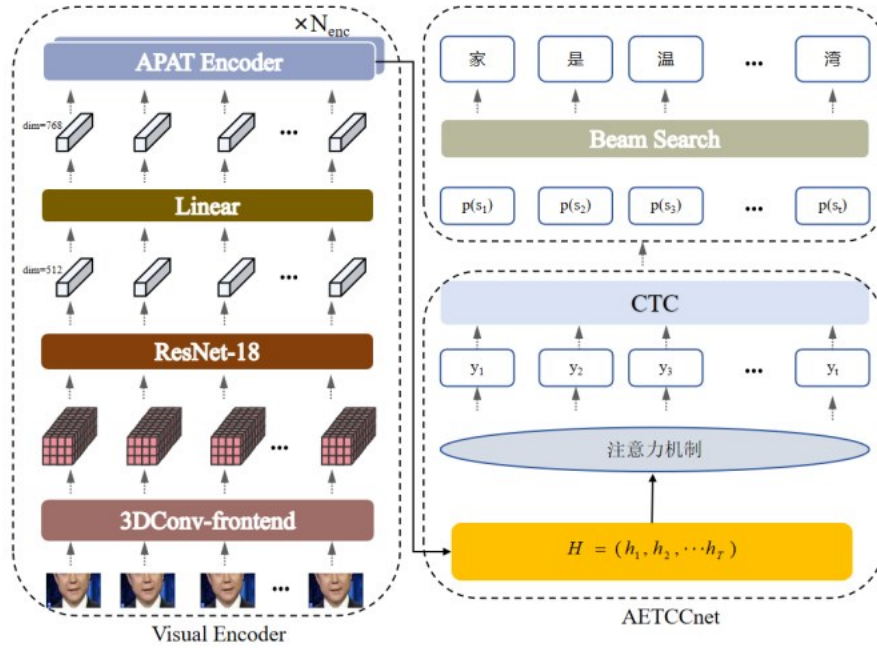


图1 APAT-LR网络结构

一个高效的特征提取和建模框架，能够精确捕捉输入数据中的局部细节及全局上下文信息。

其中自适应池化注意力模块，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倍，使用卷积核大小为 1×1 的2DCNN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 ：

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

该权重与原始特征 X 相乘，使时空级别特征进行选择强化：

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

增强后的特征送入 X' 多头自注意力模块，以捕捉全局依赖关系。首先将 X' 通过线性变换得到查询（Query）、键（Key）和值（Value）：

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层归一化，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包含一个前馈网络，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层归一化：

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

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为：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编码器利用自适应池化注意力模块提取关键帧信息，同时结合多头自注意力机制捕捉全局依赖，增强模型的时空特征。

2.3 注意力增强时序连续分类网络

连接时序分类器（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

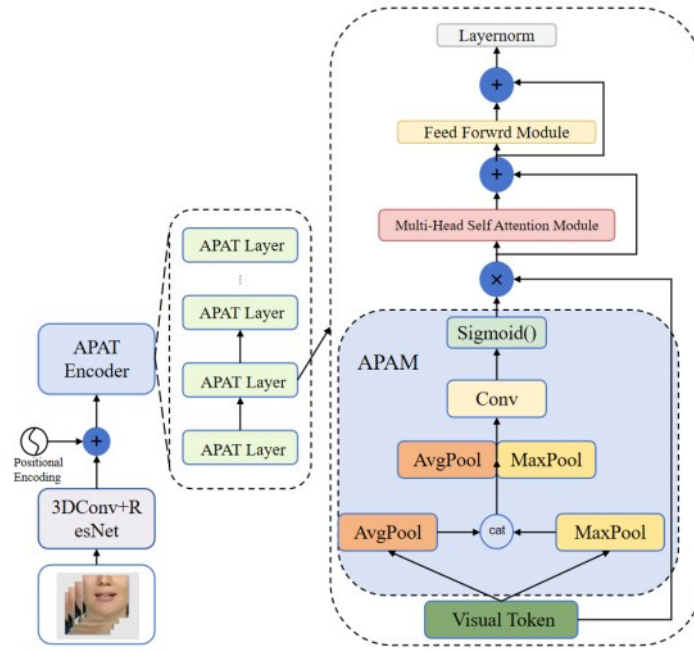


图2 APAT 编码器结构图

GRU 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 为表示在时刻 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:

$$\mathcal{L}_{CTC} = -\log P(l|x) \quad (13)$$

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:

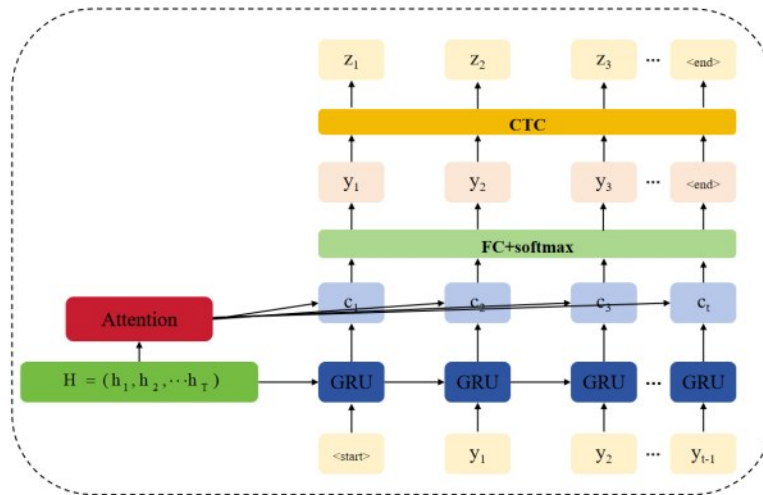


图3 AETCCnet 分类网络结构

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* 计算注意力分布:

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

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:

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

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:

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

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。

3 实验过程及其分析

3.1 数据集

本文的研究工作在 CMLR 和 GRID 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 (VIPA) 团队构建, 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 月期间中国国家新闻节目《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。数据集中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。

唇语识别性能常采用字符错误率 (Character Error Rate, CER) 作为评价指标。CER 的计算公

式如式(19)所示, 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 则为真实标签中汉字的总数。该指标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。

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

3.2 数据预处理

本文首先将视频以每秒 25 帧的速率转换为连续的图片帧。采用 *RetinaFacePredictor* 方法进行人脸检测, 精确定位人脸的大致区域。接着, 使用 *FANPredictor* 方法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 64x128 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。

3.3 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 运行内存以及 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。

3.4 数据增强

本文采用随机裁剪 (Random Cropping) 作为数据增强方法。该方法涉及在原始图像上随机选取一个坐标点, 以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 88x88 像素的图像片段。通过这种方式, 加强模型对于图像中目标对象的定位和识别能力。

3.5 训练设置

模型训练过程中采用 Adam 优化器, 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, 从而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。在模型预测阶段, 使用束搜索 (Beam Search) 进行解码,

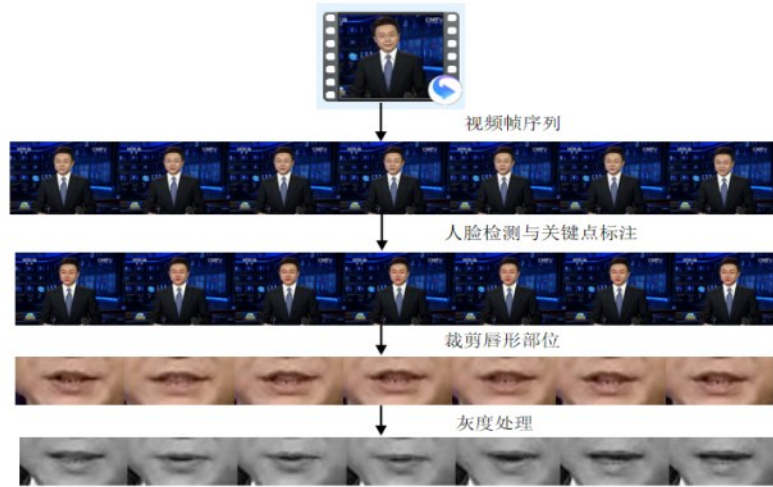


图4 数据预处理流程

以提高序列预测的准确性和鲁棒性。

3.6 与现有方法的对比

本文对比了不同方法在 CMLR 数据集和 GRID 数据集上的字符错误率, 如表 1 所示, 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 APAT-LR 方法在 CMLR 和 GRID 的测试集上的字符错误率低于其他方法, 通过使用自适应池化注意力 Transformer 编码器和注意力增强时序连续分类网络的组合在 CMLR 数据集和 GRID 测试集上获得了最佳的识别效果。

3.7 消融实验

为了验证本文所使用的数据增强方法以及方法模块的有效性, 如表 2 所示, 对数据增强方法、APAT 编码器以及 AETCCnet 进行了消融实验, 其中时序特征提使用 APAT 编码器, 分类网络采用 AETCCnet。通过消融实验结果可以观察到, 不同模块对识别准确率的提升都有帮助, 同时使用时帮助模型获得更好的识别效果。

3.8 可视化分析

为了研究提出的在处理唇形序列时的有效性, 分析了其在不同时间步长下的显著性图。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 模块模型能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。

4 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 的唇语识别算法首先, 通过 Conv3d+ResNet 视觉前端直接从原始像素中提取特征, 然后将这些特征送入 APAT-Encoder 进行时序建模。在 APAT-Encoder 的输出上, 通过 AETCCnet 解码器进行汉字预测。在模型训练过程中, 本文结合连接时序分类 (CTC) 与注意力机制, 采用混合训练策略以提升模型性能。在 CMLR 和 GRID 数据

表 1

本方法和其他方法在两个数据集的指标对比

Methods	Front-end	Back-end	CER(CMLR)	WER(GRID)
LipCH-Net ^[13]	LipCH-Net	RNN+CTC	42.1%	4.3%
WAS ^[14]	VGG	GRU(attention)	39.93%	3.1%
WLAS ^[16]	3DConv+ResNet18	LSTM(attention)	39.4%	2.7%
CSSMCM ^[17]	VGG	GRU(attention)	38.2%	2.3%
CHSLR-VF ^[18]	CNN+ResNet18	GRU+MLP	35.2%	2.4%
LIBS ^[19]	VGG-M	RNN(attention)+MLP	31.27%	2.31%
3DT-CHLipNet ^[20]	3DConv+ResNet18	TCN+CTC	30.31	2.2%
APAT-LR	3DConv+ResNet18+APAT	AETCCnet	28.4%	1.90%

表2 消融实验结果				
数据增强	APAT	AETCCnet	CMLR	GRID
			CER(%)	WER(%)
√	x	x	40.5	3.50
x	√	x	38.9	3.23
x	x	√	35.1	3.11
√	√	x	33.3	2.65
√	x	√	32.2	2.61
x	√	√	31.3	2.55
√	√	√	28.4	1.90

集上的实验结果表明, 所提出的算法在字符错误率 (CER) 这一评估指标上表现优于现有方法, 验证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。本工作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 (GAN) 等方法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。

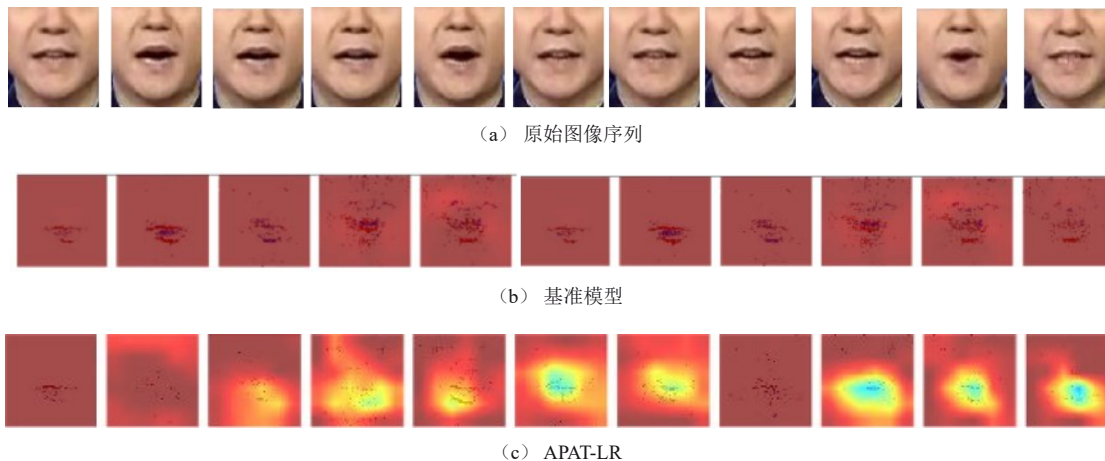


图5 唇形显著性图

参考文献:

- [1] Sumbly W H, Pollack I. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 1954, 26(2): 212-215.
- [2] Zhao Y, Xu R, Song M. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. *ACM*, 2019. DOI: 10.1145/3338533.3366579.
- [3] WENG Xinshuo and KITANI K. Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading. *Proceedings of the 30th British Machine Vision Conference*, Cardiff, UK, 2019.
- [4] Tatulli E, Hueber T. Feature extraction using multimodal convolutional neural networks for visual speech recognition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*.
- [5] Miled M, Messaoud M A B, Bouzid A. Lip reading of words with lip segmentation and deep learning[J]. *Multimedia Tools and Applications*, 2023, 82(1): 551-571.
- [6] 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(01): 73-84. DOI: 10.16451/j.cnki.issn1003-6059.202401006.
- [7] Noda K, Yamaguchi Y, Nakadai K, et al. Lipreading using convolutional neural network. *Interspeech*, 2014, 1: 3.
- [8] Xu K, Li D, Cassimatis N, et al. LCANet: End-to-end lipreading with cascaded attention-CTC[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). *IEEE*, 2018: 548-555.
- [9] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading[J]. *arXiv preprint arXiv:1703.04105*, 2017.
- [10] Zhang T, He L, Li X, et al. Efficient end-to-end sentence-level lipreading with temporal convolutional networks[J]. *Applied Sciences*, 2021, 11(15): 6975.
- [11] Ma P, Martinez B, Petridis S, et al. Towards practical lipreading with distilled and efficient models[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). *IEEE*, 2021: 7608-7612.
- [12] 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. DOI: 10.11959/j.issn.2096-6652.202252.
- [13] Assael Y M, Shillingford B, Whiteson S, et al. Lipnet: End-to-end sentence-level lipreading[J]. *arXiv preprint arXiv:1611.01599*, 2016.
- [14] Son Chung J, Senior A, Vinyals O, et al. Lip reading sentences in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6447-6456.
- [15] XUE F., YANG T., LIU K., et al. LCSNet: End-to-End Lipreading with Channel-Aware Feature Selection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2023. DOI: 10.1145/3524620.
- [16] MA P C, PETRIDIS S, PANTIC M. Visual Speech Recognition for Multiple Languages in the Wild. *Nature Machine Intelligence*. 2022, 4: 930-939.
- [17] 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]. *Journal of Computer Engineering & Applications*, 2022, 58(4).
- [18] Zhao Y, Xu R, Wang X, et al. Hearing lips: Improving lip reading by distilling speech recognizers[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 6917-6924.
- [19] 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. DOI: 10.19734/j.issn.1001-3695.2023.02.0051.
- [20] 宁佐金, 蒋近, 彭思齐. 基于标注人脸轮廓的唇语识别方法[J]. *信息技术与信息化*, 2023, (11): 199-203.
- [21] 许文稼, 李克. 基于光流与注意力机制的句级唇语识别[J]. *电子器件*, 2023, 46(05): 1339-1348.
- [22] 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.
- [23] 吴威龙, 李润恺, 许霜桦, 等. 基于自适应序列帧长度的端到端唇语识别算法[J]. *生命科学仪器*, 2023, 21(04): 35-39.
- [24] 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. DOI: 10.11959/j.issn.2096-6652.202234.
- [25] Wang H, Cui B, Yuan Q, et al. Mini-3DCvT: a lightweight lip-reading method based on 3D convolution visual transformer[J]. *The Visual Computer*, 2024: 1-13.
- [26] 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. DOI: 10.11959/j.issn.2096-6652.202128.
- [27] Wang J, Qian X, Zhang M, et al. Seeing what you said: Talking face generation guided by a lip reading expert[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14653-14662.
- [28] Prajwal K R, Afouras T, Zisserman A. Sub-word level lip reading with visual attention[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2022: 5162-5172.
- [29] Yemini Y, Shamsian A, Bracha L, et al. LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading[C]//The Twelfth International Conference on Learning Representations.
- [30] Yin Y, Wang Z, Xia K, et al. Acoustic-based lip reading for mobile devices: dataset, benchmark and a self distillation-based approach[J]. *IEEE Transactions on Mobile Computing*, 2023, 23(5): 4548-4565.
- [31] Zhang X, Zhang C, Sui J, et al. Boosting Lip Reading with a Multi-View Fusion Network[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). *IEEE Computer Society*, 2022: 1-6.
- [32] Li Z, Lohrenz T, Dunkelberg M, et al. Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention[C]//2022 IEEE Spoken Language Technology Workshop (SLT). *IEEE*, 2023: 723-730.
- [33] Xia B, Yang S, Shan S, et al. UniLip: Learning Visual-Textual Mapping with Uni-Modal Data for Lip Reading[C]//BMVC. 2023: 190-191.
- [34] Jinlin M A, Yuhao L I U, Ziping M A, et al. HSKDLR: Lightweight Lip Reading Method Based on Homogeneous Self-Knowledge Distillation[J]. *Journal of Frontiers of Computer Science & Technology*, 2023, 17(11): 2689.
- [35] He L, Ding B, Wang H, et al. An optimal 3D convolutional neural network based lipreading method[J]. *IET Image Processing*, 2022, 16(1): 113-122.
- [36] Cheng X, Jin T, Li L, et al. OpenSR: Open-Modality Speech Recognition via Maintaining Multi-Modality Alignment[C]//The 61st Annual Meeting Of The Association For Computational Linguistics. 2023.
- [37] Shi B, Hsu W N, Lakhota K, et al. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction[C]//International Conference on Learning Representations.
- [38] Burchi M, Timofte R. Audio-visual efficient conformer for robust speech recognition[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023: 2258-2267.

- [39] 洪依, 孙成立, 冷严. 基于超轻量通道注意力的端对端语音增强方法[J]. 智能科学与技术学报, 2021, 3(3): 351-358. DOI: 10.11959/j.issn.2096-6652.202136.
- [40] Peng Z, Luo Y, Shi Y, et al. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 5292-5301.
- [41] Lu J, Sisman B, Liu R, et al. Visualtts: Tts with accurate lip-speech synchronization for automatic voice over[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 8032-8036.

[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。



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。