计算机科学
COMPUTER SCIENCE

Partition-Time Masking:一种唇语识别数据增强方法

胡宇, 殷继彬

引用本文

胡宇, 殷继彬. Partition-Time Masking:一种唇语识别数据增强方法[J]. 计算机科学, 2024, 51(11A): 240300139-6.

HU Yu, YIN Jibin. Partition-Time Masking:A Data Augmentation Method for Lip Reading[J]. Computer Science, 2024, 51(11A): 240300139-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300139

Partition-Time Masking:一种唇语识别数据增强方法

胡 宇 殷继彬
昆明理工大学信息工程与自化学院 昆明 650500
(1786702137@qq.com)

摘 要 提出了一种唇语识别数据增强方法 Partition-Time Masking。该方法直接作用于输入数据,通过将输入划分为多个子序列再分别进行 Mask 操作最后再将各子序列按序拼接,使得模型能对部分帧缺失的输入具有更强的鲁棒性,从而增强泛化能力。实验前根据划分的子序列数目与掩码值来源不同而设计了 5 种增强策略,并与唇语识别研究中最重要的数据增强方法 Time Masking 进行了对比实验。实验在 LRW 数据集和 LRW1000 数据集上进行,实验结果表明 Partition-Time Masking 方法对模型性能提升的效果要优于 Time Masking 方法,其中子序列数目为 3、掩码值选择各子序列平均帧时为最优策略,该策略使得目前最佳的唇语识别模型 DC-TCN 的性能从 89.6%提高到 90.0%。

关键词:唇语识别;Time Making;数据增强;视觉语音识别;DC-TCN
中图分类号 TP391

Partition-Time Masking:A Data Augmentation Method for Lip Reading

HU Yu and YIN Jibin
Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China

Abstract This paper proposes a new data augmentation method for lip-reading called Partition-Time Masking. This method operates directly on the input data,dividing it into multiple subsequences,each undergoing a separate masking operation before being sequentially reassembled. This approach enhances the model's robustness to inputs with partial frame loss,thereby improving generalization. Five augmentation strategies are designed based on the number of divided subsequences and the source of the mask values. Comparative experiments are also conducted with the Time Masking method,a pivotal data augmentation technique in lip-reading research. Experiments are carried out on the LRW and LRW1000 datasets. The results indicate that the Partition-Time Masking method surpasses the Time Masking method in enhancing model performance. The optimal strategy is identified as using an average frame of each subsequence for masking,with the number of subsequences set to three. This approach improves the performance of the state-of-the-art lip-reading model DC-TCN from 89.6% to 90.0%.

Keywords Lip reading recognition,Time Masking,Data enhancement,Visual speech recognition,DC-TCN

1 引言

唇语识别又被称为视觉语音识别(Visual speech recognition,VSR),它是一种依赖视觉信息的语音识别技术,该技术在嘈杂环境或辅助言语障碍者交流时,展示出超越传统音频依赖语音识别的独特优势。

近年来,随着深度学习浪潮的再度兴起,唇语识别研究已从传统依赖手工特征提取和简单模式识别的方法转变为使用复杂的神经网络架构。CNN 首次用于单词识别任务时就被证明了 相较于传统特征提取的优势,实现了 38%的单词分类准确率。3DCNN 的使用极大地提升了模型对时序特征的学习能力,在 LRW 数据集上最高识别率达到了 83%。将时空卷积网络(TCN)与 ResNet34 结构相结合的模型,在 LRW 数据集上取得了 83.3%的识别准确率。受到紧密连接网络启发而设计的密集连接时序卷积网络(DC-TCN)搭配 ResNet18 网络进行特征提取,最终在 LRW 数据集上实现了 88.36%的分类准确率。

数据增强也称数据增广,指的是对原始训练数据集进行一系列变换,从而生成新的或修改过的数据样本。这一过程的主旨在于增加数据集的多样性,增强模型在遇到新的或未曾见过的数据时的泛化能力,从而提升其在实际应用中的性能和鲁棒性[1-2]。但目前唇语识别研究中所使用的数据增强方法(Mixup,CutMix 等)大多来源于其他领域,而针对唇语数据是时序数据、数据细颗粒度等特点[3],这些方法的表现低于预期。因此,适用于唇语识别领域的数据增强方法具有很大的研究前景。本文提出了一种新的唇语数据增强方法——Partition-Time Masking,该方法的核心思想是先将输入序列划分为多个子序列,再进行 Time Masking 操作,最后将各子序列按序拼接。实验结果表明该方法能提高唇语识别网络的性能和鲁棒性。

2 相关工作

数据增强方法通过增加数据集的数量和多样性,有效地解决了网络模型训练需大量数据的问题,从而有助于提升模

通信作者:殷继彬(41868028@qq.com)

型的性能[1]。目前在唇语识别研究中使用的数据增强方法种类相对较少，且多数方法源自图像处理领域。这些方法主要分为两大类：

1)未考虑数据时间维度信息的方法：这类方法都源自于图像领域，而后被应用到唇语识别领域中，操作方式与在图像处理领域类似，主要集中于对唇语数据的单帧图像进行处理，如随机裁剪、翻转、颜色调整等。Feng 等[4]提出通过引入模型先验知识进行数据增强，Zhang 等[5]则通过随机裁剪输入数据的一部分来进行增强。此外，Mixup[6]方法通过将不同的训练样本随机混合来增加样本数量，而 CutMix[7]则是通过将一个样本的一部分内容覆盖到另一个样本上来创造新的训练样本。这些方法虽然能提供视觉上的多样性，但对于模拟唇语识别中的实际挑战(例如不同的语速、口型差异、模糊不清的说话等)的能力有限。

2)考虑数据的时间维度信息的方法：这类方法源于音频或者其他时序领域，在唇语研究中使用这类方法时不仅考虑了时序数据的时间维度信息，使得时序模型能够学习到输入数据的更多时序特征，也考虑了构成时序数据的每帧所具有的视觉信息。Stafylakis 等提出了词边界(Word Boundary)[8]方法，该方法通过向模型添加含有单词边界信息的指示符作为额外的输入。指示符是与输入数据序列长度(单个视频帧数)相匹配的二进制向量，含有效信息的帧的向量值设为 1，否则设为 0。将这些单词边界信息的向量与编码器的输入拼接形成新的输入，然后送入时序模型中进行处理。Wang 等[9]基于 SpecAugment[10]提出了 Time Masking 数据增强方法，SpecAugment 方法最初在自动语音识别(ASR)领域中得到应用，后广泛用于时间序列研究。为了适应唇语识别任务，作者对该方法进行了改进，将掩码部分的值从 0 改为输入序列的平均帧，Time Masking 通过对连续的时间步进行掩码操作，提高了模型在时序序列部分缺失的情况下的鲁棒性。通过与 Mixup，Variable length augmentation[11]和 Cutout[5]等数据增强方法的对比实验，证明了 Time Masking 是目前唇语识别研究中最有效的数据增强方法。

3 Partition-Time Masking 方法

3.1 Time Masking 方法及其不足

Time Masking 方法通过对视频序列中选定的连续帧进行掩码操作来生成增强数据。其中掩码值为输入视频数据的平均帧，掩码最大长度 M 为输入帧数的 $\alpha(0.6)$ 倍，掩码的起始位置和长度是从 $[0:M]$ 随机生成得到的，如图 1 所示。

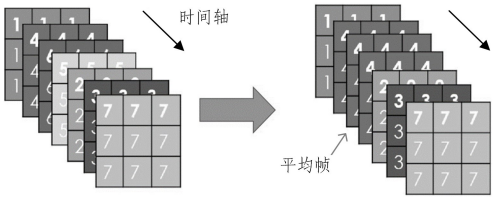

图 1 Time Masking 方法操作示意图

Fig. 1 Operation diagram of Time Masking method

Time Masking 方法不足：虽然该方法能够很好的提升模型的鲁棒性，但由于每次的掩码起始位置与长度都是随机，当使用该方法去处理的视频数据中包含相似部分的样本(唇语数据集中样本 GIVING 和 LIVING)时，由于每次掩码的起始位置和长度都是随机决定的，不同的视频片段可能在处理后仅保留相似的部分，而其他部分则被进行掩码操作，如图 2 所示。尽管掩码值采用视频数据的平均帧，但仍可能使得不同时间序列之间的相似度升高，削弱了模型在区分类似样本时的能力，而对模型分类准确率产生影响。

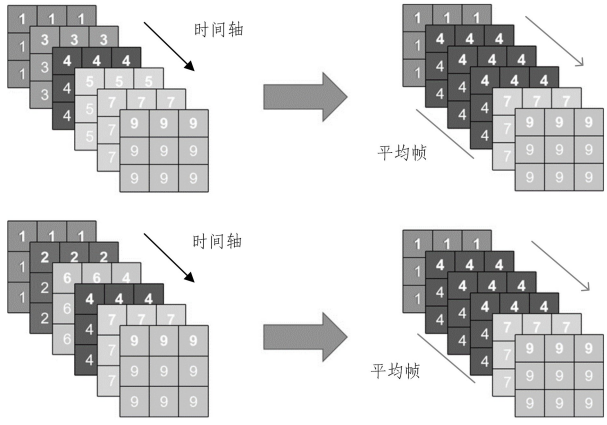

图 2 Time Masking 处理有相似部分数据示意图

Fig. 2 Processing similar data by Time Masking method

3.2 Partition-Time Masking 方法

为了改善 Time Masking 方法所存在的弊端，本文提出 Partition-Time Masking 方法。该方法核心在于：首先将输入的视频数据按序划分成多个子序列，即执行 Partition 操作；随后，对每个子序列分别进行 Time Masking 操作，其中每个子序列的掩码起始位置和长度都是独立随机生成的，以确保每个子序列掩码设置的唯一性，减少不同序列间的相似性，从而提高模型在处理相似样本时的区分能力；最终将子序列按序拼接，即得到增强后的数据。Partition-Time Masking 方法具体操作步骤如下：

1)输入数据 X，并将其划分为 k 个子序列，每个子序列包含的连续帧数为 T_k，Xn 表示第 n 个子序列；

$$T_k=\left\lfloor \frac{T}{k} \right\rfloor \tag{1}$$

$$X_n[0:T_k]=X[(n-1)T_k:nT_k] \tag{2}$$

2)计算掩码长度 t_0 值和中间变量 ∂；

$$\tau=\alpha T_k \tag{3}$$

$$\partial, t_0 \sim \bigcup(0,\tau) \tag{4}$$

3)计算掩码起始位置 t_1；

$$t_1 \sim \bigcup(0,\tau-\partial) \tag{5}$$

4)将 X_n 从 t_1 开始的连续 t_0 帧进行掩码操作；

$$X_n[t_1:t_0+t_1]=X.mean() \tag{6}$$

5)重复步骤 2－步骤 4 k 次，对每个子序列进行操作；

6)将各个子序列按序拼接，即得到增强后数据 Y；

$$Y=vstack(X_1,\cdots,X_k) \tag{7}$$

7)输出数据 Y。

为了更好地保留原始输入的时间维度信息，使用每个子序列自身的平均帧作为掩码值，而非整体输入数据的平均帧。虽然 Khan 等[8]也考虑了对数据进行多次掩码操作并进行了实验比较，在 Zhang 等[9]的 Time Masking 相关源码部分也有

实现此种想法的工作，但这些研究都只是进行重复掩码操作，即将已经进行掩码操作的数据继续进行该操作重复多次。这种重复掩码策略可能导致最终掩码长度超出设定上限，从而增加数据丢失风险。由于 Partition-Time Masking 方法是对各个子序列进行 Time Masking 操作，因此很好地避免了此类情况的出现。

3.3　Partition-Time Masking 增强策略设计

Partition-Time Masking 方法中子序列的数目和每个子序列掩码值需要提前设置而非随机生成。为了提高模型对时间维度变化的敏感性和对部分数据丢失的鲁棒性，在掩码长度不超过单次输入帧数 α 倍(0.6)的前提下，设计了一系列实验策略，以探究划分的子序列数目与子序列掩码值的来源对最终识别结果的影响。通过不同的掩码值来源 A(原始输入平均帧)、B(各子序列平均帧)分别搭配不同子序列数目(2，3，5)，而设计了 A-Mask Double(AD)，A-Mask Three(AT)，A-Mask Five(AF)，B-Mask Double(BD)，B-Mask Three(BT)这一系列的策略组合，其参数配置如表 1 所列。

表 1　增强策略参数设置

Table 1　parametersettings of enhancement policy

Policy	M	nT
AD	A	2
BD	B	2
AT	A	3
BT	B	3
AF	A	5

注：其中 nT 表示分块数目，M 表示选择的掩码策略。

4　实验准备

本章介绍了研究中主要使用的网络模型、实验的数据预处理和学习率调整策略，这些对于最终网络性能也有着重要影响。

本研究主要使用 DC-TCN 网络[9]来完成唇语识别相关研究工作，同时，为了验证 Partition-Time Masking 方法的泛化性也选择了 Bi-GRU[5]，MS-TCN[12]等唇语识别研究中的常见网络进行实验。这些都是端到端的模型，容易训练，并且由于公共唇语数据集被广泛应用，因为这些模型也大量地被应用于唇语识别研究中，相关研究文献丰富且代码易获取，这些都有利于进行本文实验，也方便与前人研究结果进行比较。

4.1　实验环境和超参数设置

实验环境为：Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz 处理器，内存 RAM 为 32 GB，显卡 GPU 选用 NVIDIA RTX3090，使用 CUDA 11.1 并行计算架构和 cuDNN 专用神经网络加速器。本文编程语言为 Python3.8.1，并使用深度学习框架 Pytorch 搭建深度学习网络。

超参数设置为：初始学习率设置为 0.000 3，总共训练 80 个 epoch，批大小为 32，使用的优化器为 AdamW，使用交叉熵[13]和 CTC[14]损失函数。

4.2　网络模型架构

DC-TCN 网络结构如图 3 所示，此网络是受紧密连接网络启发[15-17]，并将紧密连接结构应用到 TCN 网络中而探索出的新型网络[18-19]。

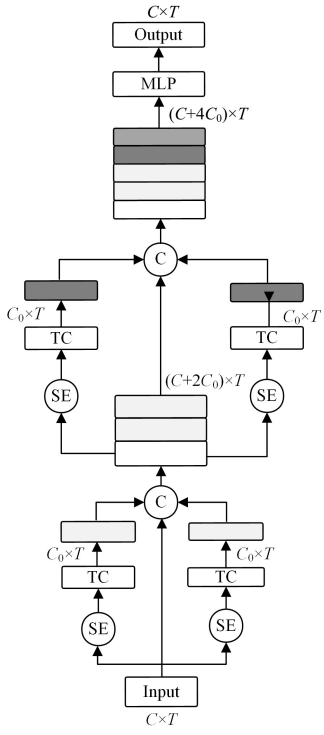

图 3　DC-TCN 网络

Fig. 3　DC-TCN network

该网络能够很好地利用浅层网络特征以有效地解决训练中出现的梯度消失问题。DC-TCN 唇语识别模型如图 4 所示。

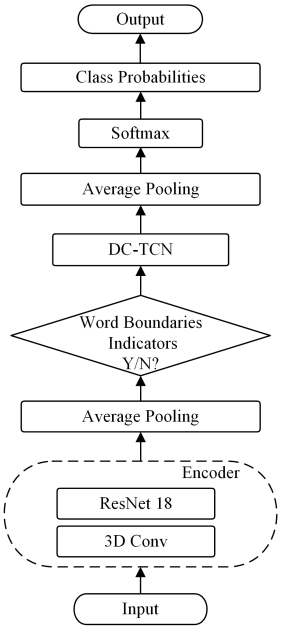

图 4　DC-TCN 唇语识别模型

Fig. 4　DC-TCN Lip reading recognition model

输入数据首先通过 3D 卷积层和 ResNet18 网络[20]组成编码器，以提取经过 Partition-Time Masking 方法处理的序列特征。然后将输出传递给解码器(时序模型)以学习相关时序特征之间的依赖性。

4.3　数据集和数据预处理

4.3.1　数据集

一个好的数据集需要满足数据量充足、种类丰富、分布均匀等条件，因此本文选择在 LRW 数据集和 LRW1000 数据集

上进行实验。

LRW 数据集是目前最大的公开可用的英文孤立词唇语识别数据集。该数据集是在 BBC 电视节目中选取了 10 00 多名访谈者的采访片段，并以短视频的形式保存，数据集中总共包含 500 个分类，每一类都为常见的单词，每个类中含有上千个样本。

该数据集由训练集、验证集和测试集 3 部分组成，其中训练集包含 488766 个样本、验证集和测试集各包含 25 000 个样本。

CAS-VAR-W1K(LRW-1000)数据集是一个自然分布的大规模中文唇语数据集[21]，用于在室外进行词级别的唇语识别和语音识别，包括 1000 个类和大约 718018 个视频样本，这些样本包含了 2000 多个说话者，总共有 100 多万个汉字实例，每一类样本包含由一个或几个汉字组成的字或者词。由于各个类别的样本数量、视频分辨率、光照条件以及说话人的姿态、年龄、性别、化妆等属性都有很大的变化，因此在该数据集上的模型训练困难。

4.3.2 数据预处理

由于不同的数据集之间存在差异性，因此数据预处理方法也有所不同，下面进行详细说明：

1)对于 LRW 数据集：需要先使用 RetinaFace 工具[22]检测视频中的人脸，然后从每个视频中截取含有人物说话部分的连续的 25 帧(约 1 秒)，其余数据丢弃。若样本有效帧不足 25 帧，则使用添加值为 0 的填充帧进行填充[8]，保证输入帧数一致。接着对每帧的图像从 $(x_1, y_1, x_2, y_2)=(80, 116, 175, 211)$ 的位置进行裁剪，以获得 96×96 的唇部图像[4]，然后将图像随机裁剪成 88×88 大小，并进行随机水平翻转后转换成灰度图像送入网络。

2)对于 LRW1000 数据集：需要先对检测视频中的人脸，截取其中所有包含人物说话内容的图片(图片数量不固定)并进行裁剪，裁剪得到大小为 128×128 的唇部图像之后缩放至 88×88 大小，再进行随机水平反转后转换为灰度图像并作为网络输入。

4.4 学习率调整策略

学习率在训练过程中的变化已被证实是影响网络性能的重要因素之一，尤其在需要对输入数据进行数据增强的情况下，本实验中采用余弦退火方法对学习率进行调整变化。同其它学习率调整方法相比，余弦退火方法在训练初期就会对学习率进行调整，并且在整个训练周期中学习率一直处于变化之中，该方法在 t 时刻下的学习率计算公式如下：

$$\eta_t=\frac{1}{2}\left(1+\cos\left(\frac{t\pi}{T}\right)\right) \tag{8}$$

其中，η 是学习率的初始值，根据先前研究经验设置为 0.003，T 是总共需要训练的轮次，在实验开始时设置为 80。与其他学习率调节策略相比，余弦退火方法从训练初期开始调整学习率，并在整个训练周期内持续进行调整。余弦退火策略的关键在于：学习率随着训练轮次的增加呈现出连续的衰减趋势，但衰减速率会逐渐减缓。学习率的这种动态变化对于网络训练的性能具有显著影响，尤其是在进行数据增强的场景中更为关键。

5 实验

本章主要介绍了常见唇语识别网络在搭配 Partition-Time Masking 方法前后在 LRW 数据集和 LRW1000 数据集上性能的变化。实验的主要目的：一是为了验证 Partition-Time Masking 方法对模型性能提升是否有效，二是对比 Partition-Time Masking 方法的不同策略对于网络性能提升的影响，从而得到最佳策略。

5.1 有效性的验证

在首次实验时采用 AD 增强策略在 LRW 和 LRW1000 数据集上对 DC-TCN，MS-TCN，Bi-GRU 这 3 个网络进行训练。Time Masking 方法可以视为子序列个数为 1 的增强策略，与 AD 增强策略仅在序列数目中有一些差别，在掩码值的选择上保持一致。AD 策略是多组策略中同 Time Masking 方法最为接近的一组，为分析 Partition-Time Masking 方法的有效性提供了重要参照。

DC-TCN 和 MS-TCN 两个网络在 LRW 数据集上进行训练，Bi-GRU 网络在 LRW1000 数据集上进行训练。此外还添加了 DC-TCN 网络搭配词边界方法的对照组，该对照组目前在 LRW 数据集上取得了最高的准确率。对比实验结果如表 2 所列。

表 2 搭配不同增强方法的模型性能

Table 2 Model performance with different enhancement methods

Model	Date Augmentation				Acc/%
	Word Boundary	TM	AD	Mix-up	
DC-TCN				√	88.01
	√	√			92.11
	√		√		91.84
		√			89.61
			√		90.03
MS-TCN				√	85.32
	√	√			88.88
	√		√		88.91
Bi-GRU	√	√			55.50
	√		√		55.70

对表 2 的实验结果进行分析，发现只将数据简单划分为两个子序列后 DC-TCN，MS-TCN，Bi-GRU 这 3 种网络模型的性能均有所提升且提升幅度要高于搭配 Time Masking 方法的对照组，DC-TCN 网络的表现最为显著，提升了 0.4%。这些实验结果验证了 Partition-Time Masking 方法的有效性，表明该方法能够有效增强模型的性能。

然而，在 DC-TCN 模型搭配词边界方法的实验组中，网络的准确率却从 92.11%下降为 91.84%。为了探究该现象出现的原因，后续研究中进行了深入的理论分析并设计了额外的补充实验以进行验证。具体实验过程后文中会详细阐明。

5.2 训练轮次的优化

在分析表 2 的实验结果时，发现 DC-TCN 网络搭配词边界是唯一个准确率下降的实验组。对该组模型的训练过程进行了深入分析，发现该组模型在训练结束前几轮准确率和损失仍然有波动，而其他模型则在接近训练结束的几轮内趋于稳定。根据相关资料和先前的研究经验，推断这可能是由于训练轮次设置过小，导致模型训练未能充分完成。为了验证这一假设，决定通过增加训练轮次来对比增加前后模型性能的变化。在前文实验中是将训练轮次设置为 80 轮，在后续实验时设置了训练轮次为 90 和 100 两个实验组，这是唇语识别研究中训练轮次常设置的数目。对 DC-TCN 模型搭配词边

界和 AD 策略、DC-TCN 模型搭配词边界和 Time Masking 两个组合重新训练，表 3 列出了不同组合在不同的训练轮次下的实验结果。

表 3　不同训练轮次下的准确率

Table 3　Accuracy rates at different training rounds

Model	Date Augmentation	*Epoch*	*Acc*/%
DC-TCN	Word Boundary+TM	80	92.11
		90	92.08
		100	91.91
	Word Boundary+AD	80	91.84
		90	92.15
		100	91.96

从表 3 中可以清楚地观察到训练轮次的改变的确会影响实验结果，当训练轮次增加到 90 时，DC-TCN 模型搭配词边界和 AD 的组合的准确率有 0.3%的提升，达到了 92.08%，并超过了搭配词边界搭配 Time Masking 的组合；当训练轮数增加到 100 时，网络准确率同 90 轮次时一样都出现了准确率降低的现象，分别降低了 0.17%，0.19%。实验结果表明关于训练轮次设置存在问题的推测是正确的，同时确定了 DC-TCN 模型搭配词边界和 AD 的组合的最佳训练轮次为 90。

5.3　不同策略的比较

前文实验结果表明了仅仅只将数据划分为两个子序列就能使得模型性能取得不错的提升。为确定 Partition-Time Masking 方法中子序列数目和掩码选择对于最终网络性能影响和寻找出最佳的策略组合，我们继续进行实验，使用 DC-TCN 网络分别搭配 BD，AT，BT，AF 增强策略在 LRW 数据集中训练。

本轮实验中训练轮次开始依旧设置为 80，如果训练结束前几轮模型的 loss、准确率还未趋于稳定则进行改变。为缩短训练周期，过程中使用 AD 策略的预训练文件进行微调[10]，实验结果如表 4 所列。

表 4　不同策略在 LRW 数据集上的性能

Table 4　Performance of different strategies on LRW dataset

Model	Policy	*Acc*/%
DC-TCN	AD	90.03
	AT	89.82
	AF	89.61
	BD	90.08
	BT	89.94

从表 4 中不难看出在搭配不同策略后模型最终性能各不相同，表明每种策略对网络性能的影响不尽相同，其中表现最好的增强策略为 BD，达到了 90.08%的准确率；其次是 AD 和 BT 策略，准确率分别为 90.03%和 89.94%，；F 策略对网络性能影响最弱，准确率为 89.61%。

数据增强的目的是为了应对训练过程中模型倾向过拟合的问题，提高模型的泛化性。从图 5 中模型在测试集中损失函数的变化曲线可以看出，由于使用了预训练模型，因此 BD 策略的 loss 在初期出现大幅度波动，而后平稳降低，并且相较于 Time Masking 组有所降低，在验证集上也表现如此。这与网络训练过程中会倾向于过拟合形成强烈反差，也证明了 Partition-TimeMasking 增强方法的有效性。

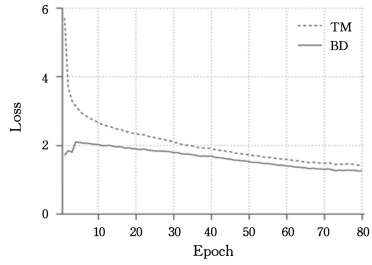

图 5　测试集上 loss

Fig. 5　Loss on test set

5.4　实验结果分析总结

从表 3 中可以看出，当训练轮次从 80 轮增加到 90 轮时，搭配 AD 策略的 DC-TCN 模型得到了更加充分的训练，从而显著提高了模型的识别准确率；而将训练轮次进一步增加到 100 时，模型可能因为过拟合而导致准确率下降，且在不同实验组中表现皆是如此。不同训练轮次下模型性能的变化充分证明了表 2 中搭配词边界和 AD 策略的 DC-TCN 模型性能降低的主要原因是由于设置的训练轮次过小，导致网络未能充分训练。因此，在对网络做出改变之后训练轮数设置也可能需要进行调整。

同时将表 2 和表 3 中实验数据进行整理与分析，能够很明显地发现当选择合适的训练轮次之后 Partition-Time Masking 方法相比于 Time Masking 方法对模型性能提升效果更显著，在 DC-TCN 搭配 BD 策略组合上提升最为明显，取得了 0.4%提升。

表 4 中列出了 DC-TCN 模型分别搭配 Partition-Time Masking 方法不同的增强策略后在 LRW 数据集上的性能。通过比较搭配各个策略后模型性能的变化，发现当子序列的数目增加到 2 和 3 时模型性能也在逐步提升，当子序列数目继续增加到 5 时模型性能降低，这可能是过拟合导致的。相较于使用原始输入的平均帧作为掩码值，选择各子序列自身的平均帧作为掩码值更能提升模型性能。目前最佳策略为 BD，即子序列数目为 3、各子序列掩码值为本序列的平均帧。

结束语　Time Masking 方法由于其掩码的起始位置和长度的随机性导致其在处理某些具有相似部分而类别不同的数据时，可能会增加这些数据的相似性，从而降低模型对这些样本的分类准确率。因而提出了 Partition-Time Masking 方法，即将输入先划分为多个子序列再进行 Time Masking 操作，为验证该方法的有效性而设计了一系列实验，得出如下结论：

1）Partition-Time Masking 方法是有效的。比较主流的唇语识别模型分别搭配 AD 策略和 Time Masking 在唇语数据集 LRW 和 LRW1000 中的表现，发现在搭配 AD 策略后绝大多数模型性能的提升幅度都要高于搭配 Time Masking 的方法。这表明了 Partition-Time Masking 方法是有效的，能进一步提高模型性能。

2）不同策略的最优训练轮次可能不同。观察表 2 中准确率降低的实验组训练日志，推测可能是由于训练轮次设置不合理而导致的。通过对该组在不同训练轮次下实验结果的比较，以验证该猜想和寻找最佳训练轮次。实验结果表明了该猜想的正确性，并得到该组的最佳训练轮次为 90。

3）BD 为最佳策略。通过比较 DC-TCN 模型搭配 Parti-

tion-Time Masking 方法不同策略的性能变化，确定了 BD 策略为最优策略，其对模型性能影响最大，取得了 0.4%提升。

以上实验结果验证了 Partition-Time Masking 方法对唇语数据识别的优越性和可行性，提高了唇语识别的准确率和鲁棒性。但本文方法仍存在以下几点问题：

1)本研究中各对照组的训练轮次和划分子序列值是非连续性选择的，有可能忽略了其他更有效的策略配置。

2)Partition-Time Masking 方法不同策略的比较实验只在单一的模型中实验，针对不同的数据集或者模型，可能不同的策略会取得更好的表现。目前由于实验周期和实验设备等客观因素的限制，在后续研究中会进行相关实验。

3)当配合词边界方法使用时，Partition-Time Masking 方法对模型性能上的提升幅度不及单独使用时的提升幅度。这可能是由于模型性能提升已经达到了瓶颈或者是这两种增强方法存在冲突，具体原因需要在未来研究中进一步验证。

参 考 文 献

[1] BAEK K, BANG D, SHIM H. GridMix: Strong regularization through local context mapping [J]. Pattern Recognition, 2021, 109: 107594.

[2] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout [J]. arXiv: 1708. 04552, 2017.

[3] XUE J, HUANG S, SONG H, et al. Fine-grained sequence-to-sequence lip reading based on self-attention and self-distillation [J]. Frontiers of Computer Science, 2023, 17(6): 176344.

[4] FENG D, YANG S, SHAN S, et al. Learn an effective lip reading model without pains [J]. arXiv: 2011. 07557, 2020.

[5] ZHANG Y, YANG S, XIAO J, et al. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition [C]// 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 356-363.

[6] WU Y, JI Q. Facial landmark detection: A literature survey [J]. International Journal of Computer Vision, 2019, 127(2): 115-142.

[7] YUN S, HAN D, OHS J, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6023-6032.

[8] STAFYLAKIS T, KHAN M H, TZIMIROPOULOSG. Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs [J]. Computer Vision and Image Understanding, 2018, 176: 22-32.

[9] MA P, WANG Y, PETRIDIS S, et al. Training strategies for improved lip-reading [C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022). IEEE, 2022: 8472-8476.

[10] PARK D S, CHAN W, ZHANG Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition [J]. arXiv: 1904. 08779, 2019.

[11] PETAJAN E D. Automatic lipreading to enhance speech recognition (speech reading) [M]. University of Illinois at Urbana-Champaign, 1984.

[12] MARTINEZ B, MA P, PETRIDIS S, et al. Lipreading using temporal convolutional networks [C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). IEEE, 2020: 6319-6323.

[13] DUDA R O, HART P E, STORK D G. Pattern classification and scene analysis [M]. New York: Wiley, 1973.

[14] MARGAM D K, ARALIKATTI R, SHARMAT, et al. LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models [J]. arXiv: 1906. 12170, 2019.

[15] GUO D, WANG S, TIAN Q, et al. Dense Temporal Convolution Network for Sign Language Translation [C]// IJCAI. 2019: 744-750.

[16] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[17] YANG M, YU K, ZHANG C, et al. Denseaspp for semantic segmentation in street scenes [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3684-3692.

[18] ZHAO X, YANG S, SHAN S, et al. Mutual information maximization for effective lip reading [C]// 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 420-427.

[19] MA P, WANG Y, SHEN J, et al. Lip-reading with densely connected temporal convolutional networks [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 2857-2866.

[20] STAFYLAKIS T, TZIMIROPOULOSG. Combining residual networks with LSTMs for lipreading [J]. arXiv: 1703. 04105, 2017.

[21] YANG S, ZHANG Y, FENG D, et al. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild [C]// 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). IEEE, 2019: 1-8.

[22] WU Y, JI Q. Facial landmark detection: A literature survey [J]. International Journal of Computer Vision, 2019, 127(2): 115-142.

HU Yu, born in 1998, postgraduate. His main research interests include deep learning and lip reading.

YIN Jibin, born in 1976, Ph.D, associate professor. His main research interests include human-computer interaction and artificial intelligence.