计算机科学
COMPUTER SCIENCE

基于改进Transformer的连续手语识别方法

王帅, 张淑军, 叶康, 郭淇

引用本文

王帅, 张淑军, 叶康, 郭淇.基于改进Transformer的连续手语识别方法[J]. 计算机科学, 2022, 49(11A): 211200198-6.

WANG Shuai, ZHANG Shu-jun, YE Kang, GUO Qi. Continuous Sign Language Recognition Method Based on Improved Transformer [J]. Computer Science, 2022, 49(11A): 211200198-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200198

基于改进 Transformer 的连续手语识别方法

王 帅 张淑军 叶 康 郭 淇
青岛科技大学信息科学技术学院 青岛 266061
(1005183361@qq.com)

摘 要 连续手语识别是一项具有挑战性的任务,当前大多数模型忽略了对长序列的整体建模能力,导致对较长手语视频的识别和翻译准确率较低。Transformer 模型独特的编解码结构可用于手语识别,但其位置编码方式以及多头自注意力机制仍有待改善。因此,文中提出了一种基于改进 Transformer 模型的连续手语识别方法,通过多处复用的带参数位置编码对连续手语句子中的每个词向量进行多次循环计算,准确掌握各个词之间的位置信息;在注意力模块中添加可学习的记忆键值对形成持久记忆模块,通过线性高维映射等比例扩大注意力头数与嵌入维度,最大程度地发挥 Transformer 模型的多头注意力机制对较长手语序列的整体建模能力,深入挖掘视频内部各帧中的关键信息。所提方法在最具权威的连续手语数据集 PHOENIX-Weather 2014[1]和 PHOENIX-Weather2014-T[2]上取得了有竞争力的识别结果。

关键词:连续手语识别;Transformer;多头注意力;位置编码
中图法分类号 TP391

Continuous Sign Language Recognition Method Based on Improved Transformer

WANG Shuai,ZHANG Shu-jun,YE Kang and GUO Qi
College of Information Science and Technology,Qingdao University of Science and Technology,Qingdao 266061,China

Abstract Continuous sign language recognition is a challenging task. Most current models ignore the overall modeling ability of long sequences,resulting in lower accuracy of recognition and translation of longer sign language videos. The unique codec structure of Transformer model can be used for sign language recognition,but its position coding method and multi-head self-attention mechanism still need to be improved. Therefore,this paper proposes a continuous sign language recognition method based on the improved Transformer model. Through multiple multiplexed position codes with parameters,each word vector in the continuous hand sentence is calculated multiple times to accurately grasp the position information between each word,add learnable memory key-value pairs to the attention module to form a persistent memory module,and expand the number of attention heads and embedding dimensions through linear high-dimensional mapping and the like,to maximize the multi-head attention mechanism of the Transformer model,and the overall modeling ability of long sign language sequences,in-depth mining of key information in each frame of the video. The proposed method achieves competitive recognition results on the most authoritative continuous sign language data sets PHOENIX-Weather2014[1] and PHOENIX-Weather2014-T[2].

Keywords Continuous sign language recognition,Transformer,Multi-head attention,Position encoding

1 引言

手语是聋人交流的主要语言,也是聋人沟通的重要媒介。一个手语动作的完成需要同时利用手动要素和非手动要素。具体来说,手动要素包括双手的形状、位置、方向和动作,而非手动要素包括眼睛、嘴形、面部表情和身体姿势。手语识别任务主要分为孤立词的识别和连续手语识别两种,后者更为贴近现实生活场景,因此成为了计算机视觉领域的主要研究对象。连续手语识别任务的目标是将手语视频自动翻译为完整的口语句子,当前大多数方法主要通过视觉信息提取、语义分割建模、文本翻译这 3 步来实现。卷积神经网络(CNN)[3]因其出色的特征提取能力,被广泛应用于视觉建模任务中;递归神经网络(RNN)[4]以及隐马尔可夫模型(HMM)[5]出色的时空建模能力使其可以很好地对简单视频序列进行有效的时序分割以及整体建模;最后,连接主义时序分类模型(CTC)[6]根据对应关系生成口语句子。然而,手语视频本身包含着大量复杂特征信息,包括单帧图像中的手部、面部、躯干之间的特征关联,以及不同帧之间各个身体部位的特征变化。因此,在处理符合真实场景下的较长手语视频的识别任务时,传统的识别方法无法在端到端的训练中提取足够多的有效特征,缺乏对视频序列的整体建模能力。

近年来,Transformer 模型[7]因其显著的长序列建模能力,从自然语言理解领域逐渐推广应用到计算机视觉领域,其全局自注意力机制的运用使序列到序列的识别和翻译任务能够并行化实现,这使得该模型成为包括连续手语识别任务等许多机器翻译任务的新架构。然而,由于 Transformer 模型

基金项目:山东省重点研发计划项目(2017GGX10127)
This work was supported by the Key Research and Development Program of Shandong(2017GGX10127).
通信作者:张淑军(zhangsj@qust.edu.cn)

位置编码能力较弱，仅使用简单的正余弦位置编码方法难以对手语视频序列中各个词向量的位置进行准确把握。另一方面，在处理连续手语识别任务时，传统的 Transformer 模型只使用简单的自注意力和多头注意力的方法，难以实现对较长手语序列的整体建模，导致识别结果较差。因此，为了使网络能更有效地提取视觉特征，提高模型的准确率和鲁棒性，本文对 Transformer 模型进行改进，使其更适合连续手语识别任务，主要提出了多处复用的含参数的位置编码、可学习的持久记忆模块以及多头注意力延伸 3 个改进模块，增强模型对长序列手语视频的整体建模能力，有效提高手语识别的准确率。

2 相关工作

早期连续手语识别方法主要依赖于人为标注的准确特征，主要使用基于图的经典方法建模。随着深度学习方法的迅速发展，2D 卷积神经网络以及 3D 卷积神经网络[8-9]凭借出色的时空表征能力被广泛应用于连续手语识别任务中。Huang 等[10]提出了采用 3D 卷积神经网络自动从原始手语视频流中提取有区别的特征进行时空建模，并与多通道视频流相融合的方法。然而，2D 卷积神经网络没有能力模拟视频序列的时间转换，3D 卷积神经网络表示状态转换的能力有限，难以准确把握视频的整体语义信息。因此，递归神经网络(RNN)被更多地用于时序建模。然而，训练 RNN 的过程中存在梯度消失以及每个时间步长(及其相关的梯度)产生的误差会逐步减小的问题。为了保持长期依赖不受梯度消失的影响，Hochreiter 等[11]提出了长短时记忆网络(LSTM)，Pigou 等[12]提出将三维残差网络与 LSTM 相结合，用于连续手语识别任务。Camgoz 等[13]提出了 SubUNets 模型，通过将专业的中间亚单元知识注入 Bi-LSTM 来实现更高的识别准确率。Xu 等[14]将 S2VT 模型用于手语识别任务，建立了 6 个张量训练模型，并将张量训练分解用于全连接层和第一个 LSTM 层，有效减少了模型参数量。然而，LSTM 循环迭代时间相对较长，且处理较长的手语序列时还会造成长距离信息的丢失。

编解码网络的提出很好地解决了以上问题，它最早出现在神经机器翻译(NMT)领域。它将源序列编码为固定大小的向量，然后从中解码目标序列，使用中间潜在空间来映射两个相关序列。Bahdanau 等[15]提出了基于 RNN 的编解码模型，并在机器翻译任务上取得了最先进的识别结果。Papastratis 等[16]提出了一种跨模态的深度学习框架，该方法由两个编码器组成，分别学习独立的视频和文本，通过线性变换将其投影到共同的潜在空间，通过联合训练的解码器对视频进行分类。

随着 NMT 领域的不断发展，许多优秀的编解码网络被提出，其中最重要的就是 Transformer 模型。2017 年，Vaswani 等[7]提出了一个基于注意力机制的 Transformer 模型，它用全注意力结构代替了递归神经网络(RNNs)，在实现并行计算的同时，使用多头注意力机制[17]对前后文之间的依赖关系进行充分捕捉，对间隔较长的词向量之间的依赖关系也能精准地把握。Transformer 不但适用于机器翻译任务[18]，也在其他各种具有挑战性的任务中取得了成功，如语言建模、句子表征学习、语音识别等。且 Transformer 模型在运算过程中充分考虑了语言翻译中的上下文语境问题，其端到端的识别方法也非常适合用于解决连续手语识别任务因此 Transformer 模型被广泛应用于连续手语识别任务中。Camgoz 等[19]是第一个将 Transformer 模型应用于连续手语识别任务中的团队，并在 RWTH-PHOENIX-Weather 2014T 连续手语数据集上取得了最优的识别结果，通过实验证明了 Transformer 模型解决连续手语识别任务的真实可行性。Niu 等[20]提出了随机帧丢弃和随机梯度停止方法，并将随机细粒度标记引入具有多状态的 Transformer 模型中，在减少视频内存占用的同时提高了模型的鲁棒性。Yin 等[21]提出了一种无需词级标注的 STMC-Transformer 模型，着重优化了连续手语任务的翻译系统。Camgoz 等[22]提出了一种新颖的多通道 Transformer 模型，结合了手形、躯干和嘴巴 3 种不同的表达形式，使模型能够从视频帧中提取到更为准确的特征信息，提高了识别准确率。Ben 等[23]提出了一个手语注意力网络 SAN，在全帧序列上对上下文序列建模，在裁剪后的手部图像上对手部序列建模，并利用自注意力机制将手部特征与其相应的时空上下文特征相结合。受以上工作的启发，本文提出了一种改进的 Transformer 模型来进行连续手语识别，有效提高了网络对较长手语序列的位置编码和整体建模能力。

3 基于改进 Transformer 模型的算法描述

本文提出的基于改进 Transformer 模型的连续手语识别方法如图 1 所示，整个算法框架由编解码器两大部分组成，模型主要针对原始方法位置编码能力不足以及长序列建模能力弱的问题，提出多处复用的可学习位置编码(Learnable Positional Embedding，LPE)、持久记忆模块(Sustainalble Memory Module，SMM)、注意力延伸模块(Attention Expansion Model，AEM) 3 个改进模块。

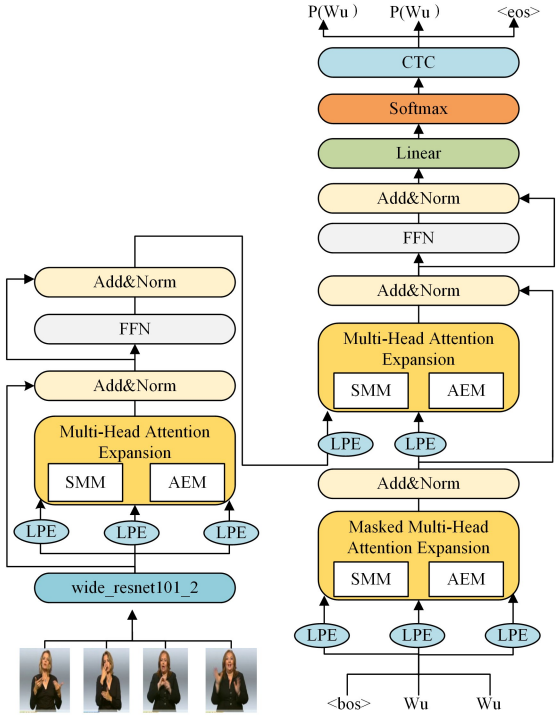

图 1 改进的 Transformer 模型整体架构

Fig. 1 Overall architecture of improved Transformer model

图 1 中，模型的总体流程如下：首先，在每个多头注意力模块前多处复用含参数的位置编码(LPE)，根据训练损失率不断更新各个字向量的位置编码权重，实现对较长句子中

各个词向量位置的精准把握；其次，针对 Transformer 模型处理长序列的手语视频时建模困难的问题，向各个多头注意力模块中加入持久记忆向量（SMM）来扩展注意力模块的深度和精度；同时，通过高维线性映射的方式同比扩大注意力模块的头数和每个头所分配的维度数（AEM），在不减少每个头所分配的感受野的前提下，进一步增强模型对较长序列的整体建模能力；最后，通过 CTC 模型对建模后的手语序列进行翻译，输出最终识别结果。通过多个模块的联合改进方法增强模型对长序列手语视频的整体建模能力，有效提高了模型对手语视频的识别准确率。

3.1　视觉特征提取

为了更好地对长序列手语视频进行建模，本文选用层数更深的 wide_resnet101_2 进行特征提取，来得到更多的手语特征信息。

不同于文本以及图像等低维度的特征提取任务，若想对连续手语视频进行多维建模，除了对视频序列使用单词嵌入外，还需要学习空间嵌入来对视频进行特征提取。本文使用 2D-CNN，从单个帧中学习提取非线性帧级空间表示，对于给定的手语视频，通过 CNN 提取空间特征的过程可表示为：

$$f_t = SpatialEmbedding(V_t) \tag{1}$$

3.2　多处复用的可学习位置编码（LPE）

由于 Transformer 模型独特的自注意力计算方式，需要对输入序列进行位置编码以防止信息丢失，一般是在编码器之前使用正余弦函数给输入序列添加位置信息。然而，较长的连续手语视频序列各个词之间的语义关系较为复杂，简单的正余弦位置编码方法难以把握长序列中手语上下文词向量间的位置关系。因此，本文采用多处复用的可学习位置编码方法，来更好地把握手语视频的整体语义信息，使词向量间的语法关系更加合理。

（1）首先，位置编码层直接继承了一个矩阵类 nn.Embedding，矩阵的长是字典的大小，宽用来表示字典中每个元素的属性向量，用于实现词与词向量的映射。

（2）其次，对 Embedding 中的权重矩阵 weight（num_words，embedding_dim）进行随机初始化，并在训练过程中不断迭代更新权重值，使模型能够自动学习更符合当前词向量的位置信息。

（3）最后，如图 1 所示，本文在每个编码器中都加入了可学习的位置编码，第一个编码器输入来自图像特征，后面的编码器输入来自前一个编码器的输出，通过多处复用的方式对手语序列中的每个关键帧实现准确编码。

3.3　加深注意力的持久记忆模块（SMM）

在面对长序列连续手语识别任务时，如何使注意力模块全面、深入、准确地挖掘到输入序列的所有特征信息，捕捉到模型的长期依赖关系是最为重要的。文献[24]提出，向多头自注意力模型中加入记忆键值向量可以实现类似前馈层的效果，从而去除前馈网络层，减少参数计算量；为了提升模型对手语序列的建模能力，本文向多头注意力模块中加入了记忆键值向量，但保留了前馈层，扩充了自注意模块的注意力深度与广度，使之更适合连续手语识别任务，本文称之为持久记忆模块（SMM）。

Transformer 模型的多头注意力层由多个并行的自注意力模块构成，它将输入特征投射到不同的子空间并在其中计算自注意力，从而捕获输入信息的多维特征。对于给定的空间向量 Y_t 以及键值对的邻接矩阵 $\boldsymbol{W}_k$ 和 $\boldsymbol{W}_v$，其键值向量为：

$$\boldsymbol{K}_t = \boldsymbol{W}_k * Y_t \tag{2}$$

$$\boldsymbol{V}_t = \boldsymbol{W}_v * Y_t \tag{3}$$

如图 2 所示，为了提高自注意力模块的注意力深度和广度，本文向 Transformer 自注意模块的前馈层向量池中添加了一组不以输入为条件的随机键值向量 $\boldsymbol{N}_k$ 和 $\boldsymbol{N}_v$，它们包含有关手语识别任务的基本常识，可以捕获不直接依赖于上下文的手语信息，并将它们在所有训练层中共享。

$$[K_1, \cdots, K_{T+N}] = Concat([W_k * Y_1, \cdots, W_k * Y_T], N_k) \tag{4}$$

$$[V_1, \cdots, V_{T+N}] = Concat([W_v * Y_1, \cdots, W_v * Y_T], N_v) \tag{5}$$

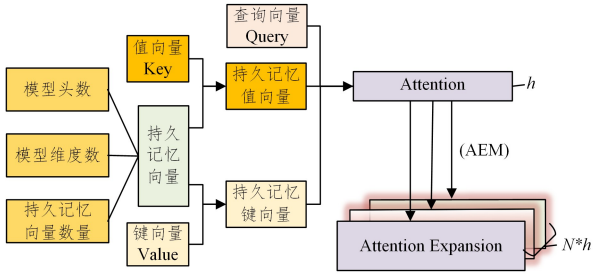

图 2　加入持久记忆向量的扩展多头注意力

Fig. 2　Extended multi-head attention with SMM

然后，自注意力层使用键向量 $\boldsymbol{K}$ 来计算输入序列的一个元素 $\boldsymbol{t}$ 与其上下文中包含了持久记忆向量的所有元素之间的相似性分数，并引入查询矩阵 $\boldsymbol{W}_q$ 进一步扩大注意力图的感受野。例如，在对手语视频建模时，词向量 $\boldsymbol{t}$ 与其上下文向量 $\boldsymbol{s}$ 之间的相似度得分可以定义为：

$$S_{t\&s} = Y_t^{\mathrm{T}} W_q^{\mathrm{T}} K_s + pos(t,s) Y_t^{\mathrm{T}} W_q^{\mathrm{T}} \tag{6}$$

其中，$\boldsymbol{pos(t,s)}$ 代表含参数的位置编码结果。对每一行使用 softmax 函数对相似度得分进行归一化，得到每个手语单词与其他词之间的注意力权重系数之和为 1 的概率分布：

$$\alpha_i = softmax(S_i) = \frac{\exp(S_i)}{\sum_{j=1}^{N} exp(S_j)} \tag{7}$$

然后，将自注意模块的头部与包含了持久记忆向量的值向量进行加权求和，输出最终的 Attention 向量：

$$Att = \sum_{i=1}^{N} \alpha_i * v_i \tag{8}$$

通过将注意力机制同时应用于输入手语词向量序列和持久记忆向量的方法，加深了注意力模块的深度与广度，在处理长序列的连续手语视频序列时，该模型可以持久地关注到相邻视频帧之间的特征变化以及每一帧内部特征间的关联信息。

3.4　注意力延伸模块（AEM）

Transformer 模型将输入序列投射到不同子空间的自注意力层来提取特征，文献[25]指出，在对 Transformer 模型进行扩展训练时，增加多头注意力模块中的头数可以提高模型性能，并增加注意力图的多样性。因此，本文将这一思想沿用至长序列的连续手语识别任务中，通过增加注意力头数来增强模型的注意力深度，使模型更好地关注到序列的整体特征信息。然而，对于具有固定嵌入维度的模型来说，直接简单地增加注意力头的数量会减少分配给每个头的维度，而维度的减少同样会影响注意力图的多样性。为了解决这一问题，

本文并没有采用直接扩展的方法，而是通过线性变换矩阵 $W_A \in R^{T*T'}$ ($T'>T$) 把注意力图 $A=[A^1,\cdots,A^H]$ 映射到 $\widetilde{\boldsymbol{A}}=[\widetilde{\boldsymbol{A}}^1,\cdots,\widetilde{\boldsymbol{A}}^{H'}]$，并满足：

$$\widetilde{\boldsymbol{A}}^h=\sum_{i=1}^{H}W_A(h,i)*A^i,h=1,\cdots,H' \tag{9}$$

该方法将多头自注意力模型线性映射到高维空间中，在适当增加注意力头数量的同时，也能够保证每个头的维度数是不变的，使模型既能享受到更多注意力头的好处，又能享受到高嵌入维度的优势。

3.5 基于 CTC 的序列对齐

由于端到端的识别任务很难获得强注释，为了能够更好地学习两个序列之间的映射关系，以端到端的方式从弱注释数据中进行训练，Graves 等提出了连接主义时间分类模型(Connectionist Temporal Classification，CTC)，它在计算误差时考虑了两个序列之间所有可能的对齐方式。因此，本文采用 CTC 对前述改进的 Transformer 模型获取到的序列进行对齐和规整，以优化最终的手语翻译结果。

当使用通用损失函数训练具有 $\boldsymbol{L}$ 个词汇表的网络时，本文将网络构建为 $\boldsymbol{L}$ 个输出，每个输出对应一个标签。CTC 引入了一个空白标签 $\boldsymbol{b}$ 并创建了一个扩展词汇表 $\boldsymbol{L}'$，其中 $\boldsymbol{L}'=\boldsymbol{L}\cup\{\boldsymbol{b}\}$，并添加另一个与空白标签对应的输出单元来重构网络。空白标签考虑了序列中目标标签之间可能存在的关联性，从而消除了对每帧注释的需要。最后 CTC 通过折叠重复去除输出结果中的空白标签和重复冗余字，输出手语序列的翻译结果。

4 实验结果与分析

本节阐述实验所用的数据集以及评价模型性能的指标、实验细节以及识别结果与分析。

4.1 数据集

(1)PHOENIX-Weather2014 是亚琛工业大学提出的德国手语数据集，语料库包括 9 个手语者打出的 7000 个天气预报句子。该数据集由 RGB 相机以每秒 25 帧的速度采集，总计 963×10^3 帧，单帧图像分辨率为 210 * 260，总词汇量为 1081。实验过程中，5672 个实例用于训练，540 个实例用于验证，629 个实例用于测试。

(2) PHOENIX-Weather2014-T 数据集是 PHOENIX-Weather2014 数据集的扩展，它专为手语翻译任务而设计，并被广泛用于评估连续手语识别任务。该数据集同样由 9 个手语者录制完成，总词汇量为 1085。实验过程中，训练、验证、测试集分别有 7096，519 和 642 个实例。

以上两个数据集为当前国内外研究手语视频识别的经典 benchmark。

4.2 评价指标

本文用当前连续手语识别任务最常用的误字率 Word Error Rate(WER)作为评价指标，WER 定义为将模型识别结果转换为正确答案需要进行的替换(sub)、插入(ins)和删除(del)操作的最小总和，WER 值越低，模型效果越好，准确率越高。同时记录每次实验的 del/ins 值作为评价模型的辅助参考。

$$WER=\frac{\#substitutions+\#insertions+\#deletions}{\#glossesinreference} \tag{10}$$

4.3 实验环境配置和参数说明

本文实验环境如表 1 所列。

表 1 实验环境配置信息

Table 1 Configuration information of experimental environment

python3.6	annaconda3	cuda10.0
tensorflow1.14.0	pytorch1.7.0	Torchvision=0.8.0
linux 4.15.0	内存：16 GB DDR4-2400 MHz * 16	GPU：NVIDIA GTX 1080Ti * 4

本文采用 Adam 优化器来训练网络，每个模型都在数据批次大小为 8 的前提下训练了 30 轮，学习率为 10^{-3} ($\beta_1=0.9$，$\beta_2=0.998$)，选用去除全连接层的 wide_resnet101_2 作为 CNN 模型提取特征；Transformer 模型头数设置为 16，维度为 1024，输入嵌入和最终表示值均为 0.1，Dropout 层参数设置为 0.3，前馈层维度数为 2048，并向 18 个自注意层中加入 2048 个持久记忆向量。

本文对所有持久记忆向量按照原始向量的维度数和阈值重新参数化，并在所有头部共享位置嵌入，使添加的持久记忆向量具有与原始上下文向量相同的单位方差。

4.4 实验结果及分析

本文方法及对比方法在 PHOENIX-Weather2014 数据集以及 PHOENIX-Weather2014-T 数据集上的实验结果对比如表 2、表 3 所列。

表 2 本文方法与对比方法在 PHOENIX-Weather2014 数据集的最新实验结果对比

Table 2 The latest experimental results comparison of method in this paper and comparison method on PHOENIX-Weather2014 dataset

实验方法	Del/ins	WER/%
DeepSign[26]	—	38.8
Re-sign[27]	—	26.8
SubUNets[13]	—	40.7
Staged-Opt[28]	—	38.7
LS-HAN[29]	—	38.3
Align-iOpt[30]	13.0/2.5	36.7
SF-Net[31]	—	34.9
DPD+TEM[32]	9.3/3.1	34.5
SFD+SGS[20]	9.3/6.6	29.4
CNN-LSTM-HMM[33]	—	26.0
SFD+SGS+SFL+LM[20]	10.4/3.6	25.8
SAN[23]	—	29.7
LPE+SMM+AEM(本文方法)	**8.5/4.7**	**25.1**

表 3 本文方法与对比方法在 PHOENIX-Weather2014-T 数据集上的最新实验结果对比

Table 3 The latest experimental results comparison of method in this paper and comparison method on PHOENIX-Weather2014-T dataset

实验方法	注释				WER/%
	词级标注	嘴部	手部	文本	
CNN-LSTM-HMM (1-Stream)[33]	√				26.5
CNN-LSTM-HMM (3-Stream)[33]	√	√	√		24.1
SLT[19]	√				24.6
SLT(Gloss+Text)[19]	√			√	24.5
SFD+SGS[20]	√				26.8
SFD+SGS+SFL[20]	√				26.1
LPE+SMM+AEM (本文方法)	√				**24.3**

由表 2、表 3 可见，本文针对 Transformer 改进的实验方法在 PHOENIX-Weather2014 数据集上得到了误字率降低 1.9%的 SOTA 识别结果，在 PHOENIX-Weather2014-T 数据集上取得了有竞争力的识别结果。

图 3 和图 4 给出了本文模型在 PHOENIX-Weather2014 数据集上使用 Baseline 方法以及改进后方法的对比，可见改进后的模型 Loss 曲线更平滑且收敛更快，最终的 Loss 值更低。

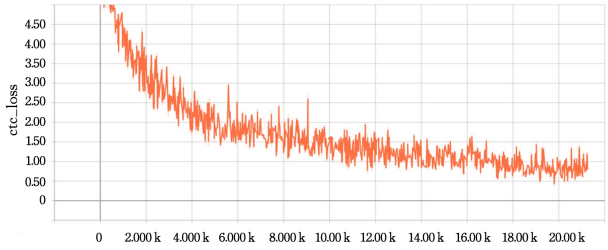

图 3　Baseline 方法在 PHOENIX-Weather2014 数据集上的 Loss 曲线

Fig. 3　Loss curve of Baseline on PHOENIX-Weather2014

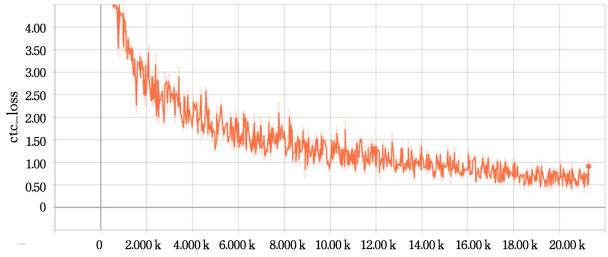

图 4　本文方法在 PHOENIX-Weather2014 数据集上的 Loss 曲线

Fig. 4　Loss curve of the proposed method on PHOENIX-Weather2014

为了验证网络结构中选用不同模块对整体性能的影响，本文在 PHOENIX-Weather2014 数据集上进行的消融实验结果如表 4—表 6 所列。

表 4　不同类型 CNN 效果对比

Table 4　Comparison of different types of CNN

（单位：%）

实验方法	WER
Resnet18(Baseline)	27.0
ResNet34	27.3
ResNet50	27.1
Resnet152	26.9
Wide_ResNet50_2	26.9
Wide_ResNet101_2	**26.8**

由表 4 可以看出，wide_resnet101_2 由于具有较宽的特征图、较大通道数以及较深的层数，可以更准确地提取手语视频的特征，取得了最优的识别结果，因此本文选用 wide_resnet101_2 模型作为手语视频的特征提取网络。

表 5　是否同步扩展注意力头数与嵌入维度效果对比

Table 5　Comparison of whether or not synchronously expanding the number of attention and embedded dimension

注意力头扩展倍数	是否同比增加嵌入维度	WER/%
1(Baseline)	是	27.0
2	是	27.1
3	是	27.0
4	是	**26.8**
5	是	26.9
6	是	28.6
4	否	27.3
5	否	27.2

由表 5 可知，当注意力头数扩展为原来的 4 倍（16 个）时，模型的注意力机制可以发挥到最优。如果继续增加模型的注意力头数，模型会出现梯度爆炸，反而导致识别准确率的降低。实验结果表明，同比增加嵌入维度对模型的识别性能有积极效果。

表 6　不同模块 LPE，SMM，AEM 的消融实验结果

Table 6　Ablation experimental results of LPE，SMM and AEM in different modules

实验方法	Del/ins	WER/%
Baseline(Transformer)	7.5/6.3	27.0
LPE	8.0/6.5	26.8
SMM	8.6/4.9	25.8
AEM	7.0/6.2	26.7
LPE+SMM	10.5/3.6	25.5
LPE+AEM	9.4/4.5	26.7
SMM+AEM	8.5/4.1	25.3
LPE+SMM+AEM(本文方法)	**8.5/4.7**	**25.1**

由表 6 可知，LPE 模块通过多处复用可学习位置编码的方式处理连续手语视频，实现了对连续手语视频的准确位置编码，并将模型的误字率降低了 0.2%；AEM 模块同样被证明可以有效扩展注意力的深度与广度，在处理连续手语视频这一长序列任务时，注意力头数和嵌入维度的增大有效降低了 0.3%的误字率；而 SMM 模块因其独特的持久记忆机制显著降低了 1.2%的模型误字率，进一步证明了注意力机制在处理连续手语识别等长序列任务时的重要性，持久记忆模块的加入增强了模型的整体建模能力。

结束语　本文提出了一种基于改进 Transformer 模型的连续手语识别方法，针对连续手语视频，使用 wide_resnet101_2 进行视觉特征提取，针对传统 Transformer 模型位置编码能力弱，难以对长序列手语视频进行建模的问题，通过多处复用的可学习位置编码、加深注意力的持久记忆模块、针对长序列的注意力延伸模块对 Transformer 模型进行优化和改进，最后通过连接主义时序分类模型进行文本翻译，最终在 PHOENIX-Weather2014 数据集上取得了当前最优的识别结果。未来工作将会探索多模态数据的整合方式，例如将手部、光流和骨骼关节点同时送入网络进行训练，使模型从连续手语视频中学习到更为准确鲁棒的特征信息，进一步提高了模型的准确率。

参 考 文 献

[1]　FORSTER J，SCHMIDT C，HOYOUX T，et al. RWTH-PHOENIX-Weather：A Large Vocabulary Sign Language Recognition and Translation Corpus[C]// International Conference on Language Resources and Evaluation(LREC). 2012.

[2]　FORSTER J，SCHMIDT C，KOLLER O，et al. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-Weather[C]// International Conference on Language Resources and Evaluation. 2014：1911-1916.

[3]　LECUN Y，BOTTOU L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[4]　LIPTON Z C，BERKOWITZ J，ELKAN C. A critical review of recurrent neural networks for sequence learning[J]. arXiv：1506.00019，2015.

[5]　FORSTER J，KOLLER O，OBERDÖRFER C，et al. Improving

Continuous Sign Language Recognition: Speech Recognition Techniques and System Design[C]// Workshop on Speech and Language Processing for Assistive Technologies. 2013.
[6] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// Proceedings of the 23rd International Conference on Machine Learning. 2006: 369-376.
[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.
[8] CUI R, LIU H, ZHANG C. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7361-7369.
[9] MOLCHANOV P, YANG X, GUPTA S, et al. Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4207-4215.
[10] JIE H, ZHOU W, LI H, et al. Sign Language Recognition using 3D convolutional neural networks[C]// IEEE International Conference on Multimedia and Expo. 2015: 1-6.
[11] HOCHREITERS, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[12] PIGOU L, HERREWEGHE M V, AMBRE J D. Gesture and Sign Language Recognition with Temporal Residual Networks [C] // IEEE International Conference on Computer Vision Workshop. 2017: 3086-3093.
[13] CAMGOZ N C, HADFIELD S, KOLLER O, et al. SubUNets: End-to-End Hand Shape and continuous sign language recognition[C]// IEEE International Conference on Computer Vision. 2017: 3075-3084.
[14] XU B, HUANG S, YE Z. Application of Tensor Train Decomposition in S2VT Model for Sign Language Recognition[J]. IEEE Access, 2021, 9: 35646-35653.
[15] BAHDANAUD, CHO K, BENGIO Y. Neural machine translation by Jointly Learning to align and translate[J]. arXiv: 1409.0473, 2014.
[16] PAPASTRATIS I, DIMITROPOULOS K, KONSTANTINIDIS D, et al. Continuous Sign Language Recognition Through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space[J]. IEEE Access, 2020, 8: 91170-91180.
[17] LIN Z, FENG M, SANTOS C, et al. A Structured Self-attentive Sentence Embedding [J]. arXiv: 1703.03130, 2017.
[18] GEHRING J, AULI M, GRANGIER D, et al. Convolutional Sequence to Sequence Learning[C]// International Conference on Machine Learning. PMLR, 2017: 1243-1252.
[19] CAMGOZ N C, KOLLER O, HADFIELD S, et al. Sign language transformers: Joint end-to-end sign language recognition and translation[J]. arXiv: 2003.13830, 2020.
[20] NIU Z, MAK B. Stochastic Fine-Grained Labeling of Multi-state Sign Glosses for Continuous Sign Language Recognition[M]. Springer, Cham, 2020.
[21] YIN K, READ J. Better Sign Language Translation with STMC-Transformer[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020.
[22] CAMGOZ N C, KOLLER O, HADFIELD S, et al. Multi-channel Transformers for Multi-articulatory Sign Language Translation [C]// European Conference on Computer Vision. Cham: Springer, 2020: 301-319.
[23] BEN SLIMANE F, BOUGUESSA M. Context Matters: Self-Attention for Sign Language Recognition[J]. arXiv: 2101.04632, 2021.
[24] SUKHBAATAR S, GRAVE E, LAMPLE G, et al. Augmenting Self-attention with Persistent Memory[J]. arXiv: 1907.01470, 2019.
[25] TOUVRON H, CORD M, SABLAYROLLES A, et al. Going deeper with Image Transformers[J]. arXiv: 2103.17239, 2021.
[26] KOLLER O, ZARGARANO, NEY H, et al. Deep sign: hybrid cnn-hmm for continuous sign language recognition[C]// British Machine Vision Conference(BMVC). 2016: 1-12.
[27] KOLLER O, ZARGARAN S, NEY H. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4297-4305.
[28] CUI R, LIU H, ZHANG C. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7361-7369.
[29] HUANG J, ZHOU W, ZHANG Q, et al. Video-based sign language recognition without temporal segmentation[C]// AAAI Conference on Artificial Intelligence. 2018.
[30] PU J, ZHOU W, LI H. Iterative alignment network for continuous sign language recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2019.
[31] YANG Z, SHI Z, SHEN X, et al. SF-Net: Structured feature network for continuous sign language recognition[J]. arXiv: 1908.01341, 2019.
[32] ZHOU H, ZHOU W, LI H. Dynamic pseudo label decoding for continuous sign language recognition[C]// IEEE International Conference on Multimedia and Expo. 2019: 1282-1287.
[33] KOLLER O, CAMGOZ C, NEY H, et al. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(9): 2306-2320.

WANG Shuai, born in 1997, postgraduate. His main research interests include computer vision.

ZHANG Shu-jun, born in 1980, associate professor. Her main research interests include computer vision.