Transformer-Based Acoustic Modeling For Hybrid Speech Recognition

基于变压器的混合语音识别声学模型

*王永强*1*阿卜杜拉赫曼·穆罕默德*1*，公爵*1*，刘春喜*1*，亚历克斯肖*1*,*

*杰伊·马哈迪奥卡尔*1,?*，黄红照*1,?*安德罗斯·特詹德拉*2,?†*，张晓辉*1,?*，弗兰克·张*1,?*，克里斯蒂安·富根*1,?*，杰弗里·茨威格*1,?*迈克尔·L·萨尔茨*1,?1              2

                                      Facebook AI，美国奈良科学技术研究所，日本

摘要

We propose and evaluate transformer-based acoustic models (AMs) for hybrid speech recognition. Several modeling choices are discussed in this work, including various positional embedding methods and an iterated loss to enable training deep transformers. We also present a preliminary study of using limited right context in transformer models, which makes it possible for streaming applications. We demonstrate that on the widely used Librispeech benchmark, our transformer-based AM outperforms the best published hybrid result by 19% to 26% relative when the standard n-gram language model (LM) is used. Combined with neural network LM for rescoring, our proposed approach achieves state-of-the-art results on Librispeech. Our findings are also confirmed on a much larger internal dataset.

Index Terms— hybrid speech recognition, acoustic modeling, transformer, recurrent neural networks

提出并评价了基于变压器的混合语音识别声学模型。在这项工作中讨论了几种建模方法，包括各种位置嵌入方法和一种迭代损失，以便能够训练深度变换器。我们还初步研究了在transformer模型中使用有限的右上下文，这使得流应用成为可能。我们证明，在广泛使用的Librispeech基准上，使用标准gram语言模型（LM）时，基于transformer的AM的性能要比最好的混合结果高出19%到26%。结合神经网络LM进行重排序，在Librispeech上取得了最新的结果。我们的发现也在一个更大的内部数据集上得到了证实。*不*

*索引项*-混合语音识别，声学建模，变压器，递归神经网络

2019.10.22 – 2020.04.30. arxiv, icassp 2020

# 1        介绍

自从在自动语音识别（ASR）中引入深度学习（deep learning）[1]以来，人们探索了各种用于声学建模的神经网络结构[2–6]。其中，递归神经网络（RNN），特别是长-短期记忆（LSTM）[7]神经网络被广泛应用于传统的混合系统（如[3，8]）、基于序列的系统（如[9，10]）或基于神经传感器的端到端系统（如[11]）。然而，RNNs有几个众所周知的局限性：1）由于[12]中发现的消失或爆炸梯度问题，RNNs不能很好地模拟长期的时间依赖性；2）RNNs的递归性使得并行处理语音信号变得困难。为了解决这些问题，人们提出了多种神经网络结构来代替RNN，包括时延神经网络（TDNN）[5]、前馈顺序记忆网络（FSMN）[6]和卷积神经网络（CNN）[4,13]，但取得的成功有限。

最近，自我注意网络[14]在各种自然语言处理任务（例如[14–16]）中显示出了有希望的结果。与RNN和CNN不同，自我注意直接连接输入序列中的任意位置对。要在两个输入位置之间向前（或向后）发送信号，只需一步就可以通过网络，而在RNNs和CNNs中只需（n）步和（logn）步。此外，自我注意的计算可以很容易地并行化。在自我注意的基础上，transformer模型[14]利用了多头注意和前馈层的交错。ASR也使用了自我注意和变压器模型，主要是在sequenceto-sequence架构中[17–19]，但值得注意的例外是[20,21]。在这项工作中，我们提出并评估基于变压器的acous-*不OO*

|  |
| --- |
| *?*平等贡献；  †Andros在Facebook实习时完成了这项工作。 |

混合ASR的tic模型（AMs）。我们探讨了几种建模方法，包括将绝对或相对位置信息编码到变压器输入中的方法，以及使深度变压器训练成为可能的迭代损耗。虽然我们在这项工作的重点是调查潜在的变压器为基础的AMs没有任何限制，我们探讨了可流动变压器和目前我们的初步实验结果。结果表明，在广泛使用的Librispeech基准和我们的内部数据集上，我们提出的基于变压器的AMs可以在非常强的双向LSTM（BLSTM）基线上产生显著的字错误率（WER）改进。当使用标准的4-gram语言模型（LM）时，我们在Librispeech上获得的结果比以前最好的混合模型提高了19%到26%；结合神经LM重构，我们的系统在这个数据集上达到了最先进的性能。

# 2        混合体系结构

在混合ASR[22]中，使用声学编码器将输入序列x·····编码为高级嵌入向量z···的序列，。这些嵌入向量用于产生隐马尔可夫模型（HMM）的绑定状态的后验分布，如senone[23]或chenone[24]。然后将这些后验分布与其他知识源（如词典和LMs）相结合，构造一个搜索图。然后使用解码器来找到最佳假设。不同的神经网络可以用作编码器：在DNN、TDNN和CNN中，zi是x的函数及其固定数量的相邻帧；在单向RNNs中，zi是xto x的函数，而在双向RNNs中，zi是整个输入序列的函数。1*,十T型*1*,z轴T型t型t型t型*1 *t型t型*

尽管与序列到序列或神经传感器结构相比，混合方法不太吸引人，因为它不是端到端的训练，但对于作者的实际问题，它仍然是性能最好的系统。它还有一个优点，即它可以很容易地与培训期间可能无法获得的其他知识源（如个性化词典）集成。在这项工作中，我们的目标是利用变压器来改善混合声学模型。

# 3        基于变压器的声学建模

在本节中，我们首先简要回顾变压器网络，并讨论将变压器用作声学编码器时的各种建模选择。第3.5节还讨论了与其他工程的关系。

3.1自我注意和多头注意

自我注意首先使用点积注意计算输入序列上的注意分布，即对于每个x∈R，分布α通过以下公式获得：*t型d级我t型*

实验（β·TqTk）*十t型***栈单***十τ*



*αtτ*=P exp（β·*十*T型*t型***栈单**问T型k公司*十τ*0)(1)

*τ*0

其中∈r变换x**栈单**问*,*k公司*d级k公司*×天*我t型*查询和键空间，是一个比例因子。注意，对于语言建模，当前位置和未来位置之间的点积被屏蔽为—∞以防止将来的信息泄漏到当前嵌入。虽然对于声学建模来说，可以关注整个序列，但是在许多应用中，我们只关注有限的右上下文帧来实现语音信号的流处理（即，+R之间的点积被屏蔽为−∞）。给定α，自我注意的输出嵌入通过以下方式获得：*t型τ、 τ>tt型*

*z轴t型*=十衰减（αtτ）·Wv*十τ*(2)

*τ*

其中∈Rmaps将输入向量映射到值空间。**W型**五*d级五*×天*我*

自我注意通常与多头部注意（MHA）相结合，其中自我注意头部被单独应用于输入序列，每个头部的输出被串联并线性转换为公共空间，即。，*小时*

                                ...

*z轴t型*=Wo退出（（3）第*τ*

其中∈R，和**栈单**o*d级我*×高清五*αtτ*（一）是注意力权重和第-个头的值矩阵。*我*

3.2变压器结构

除了MHA子层外，每个变压器层还包含一个完全连接的前馈网络（FFN），该网络由两个线性变换和一个非线性激活函数组成。将FFN网络分别相同地应用于序列中的每个位置。为了允许将多个变压器层堆叠在一起，将剩余连接添加到MHA和FFN子层。作为正则化的一种形式，在MHA和线性变换之后也应用了辍学。图1总结了一个转换器层的体系结构。注意，与[14]不同，在MHA和FFN之前应用层归一化[25]，并且第三层归一化（图1中的LN3）是必要的，以防止完全绕过变压器层。注意，在[15]之后，我们在FFN网络中使用“gelu”非线性[26]。

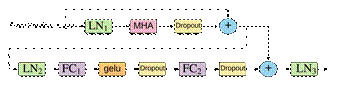


图1：一个变压器层的结构。“LN”表示层归一化[25]；“FC”表示全连通线性变换；“gelu”表示gelu非线性激活[26]。

3.3位置嵌入

变换器层的一个明显的限制是输出对输入顺序排列是不变的，即对于任何应用于输入序列x············································，。这意味着transformer不为*π*1*,十T型π*1*,z轴T型*

输入序列的顺序。在最初的变压器工作[14]中，通过正弦位置嵌入将绝对位置的信息注入到输入序列中来解决这一问题。我们认为，与自然语言处理不同，相对位置对语音信号更有用。在这项工作中，我们比较了几种将位置信息编码到变压器输入中的方法：

•向x添加正弦位置嵌入pis，其中pis的第n个元素sin（（t/10000）i/d）表示偶数，cos（（t/10000）（i−1）/d表示奇数。它编码绝对位置信息；*正弦位置嵌入：t型t型我t型我我我我*

•*帧堆叠：*打破排列不变性的一个简单方法是将上下文向量堆叠在一起，即x=*不t型*

T型

. 这是相对位置的编码-

*民族的*信息；

•受[27]的启发，我们使用二维卷积层对相对位置信息进行隐式编码。卷积嵌入隐式地执行帧叠加，并学习有用的短程频谱时域模式[28]。*卷积嵌入：*

3.4      训练深度变形金刚

变压器层可以堆叠多次，形成一个非常深的网络。在我们最初的实验中，我们发现更深的网络可以获得更好的精度。然而，在叠加许多层之后，训练变得困难，并且经常陷入一个糟糕的局部最优。为了训练深度变压器，我们使用了迭代损耗[29]，其中一些中间变压器层的输出也用于计算辅助交叉熵损耗。对这些辅助损耗进行插值，得到最终的损耗函数。注意，中间层特定参数（例如，softmax操作之前的线性变换）在训练之后被丢弃。

3.5      与其他工程的关系

原始的transformer论文[14]提出用自我注意和交叉注意来代替序列到序列模型中编码器和解码器中的递归。由于我们主要研究混合语音识别，所以在本研究中，我们仅使用自我注意来取代声学编码器中的RNN。

基于自我注意的声学建模在过去已经被探索过。在[20]中，自我注意被修改为关注固定数量的左右上下文框架，并且只使用一个注意层。相比之下，在我们的工作中，注意头注意到所有过去的帧，并且我们使用具有非常深结构的自我注意和FFN网络，这对于获得良好的模型精度是至关重要的。在[30]中，变压器与RNN在顺序到顺序架构中进行了比较。在文献[18]中，研究了序列到序列模型的各种位置嵌入方法，发现用LSTM层代替FFN网络，使自我注意层具有位置感知能力，可以获得更好的性能。在[27]之后，我们使用卷积层作为变压器层输入的预处理器，并将其与第4.2节中的其他位置编码方法进行比较。在[31]中，类似于迭代损失的损失函数用于训练字符级LMs的深度变换器；我们证明了它对于训练基于深度变换器的AMs也是至关重要的。

# 4        实验

为了评估所提出的基于变压器的声学模型的有效性，我们首先在Librispeech语料库上进行了实验[32]。这个语料库包含大约960个小时的语音数据，用于训练，以及4个开发和测试集（{dev，test}-{clean，other}），其中其他集更具声学挑战性。不对这些测试集执行分段。标准的4-gram语言模型（LM）具有200K词汇表，用于所有的一次通过解码。

4.1实验装置

在[24]之后，我们在所有实验中都使用了上下文和位置相关的图形（即chenones）。我们使用标准的Kaldi[33]Librispeech配方引导HMM-GMM系统。我们使用具有固定自循环和前向转移概率（均为0.5）的1状态HMM拓扑。采用10ms的帧移提取80维log-Mel滤波器组特征。降低的20ms帧速率可通过堆叠和跨步2个连续帧或通过卷积层中的跨步2池（如果使用）来实现。我们发现，这不仅减少了计算量，而且略微提高了识别精度。使用了速度扰动[34]和SpecAugment[10]（没有时间扭曲的LD策略）。我们专注于交叉熵（CE）训练模型，仅在最佳CE设置的基础上选择性地执行sMBR[35]训练。

使用基于PyTorch的fairseq[36]工具箱内部开发的语音扩展来执行神经网络训练。Adam optimizer[37]用于所有实验；在前8000次迭代中，学习率线性地从1e-5上升到1e-3，在其余的训练中保持在1e-3。在这项工作中，我们主要比较了全上下文转换器和BLSTM，尽管我们对使用有限右上下文的转换器进行了初步的研究。所有实验中都使用了压降：变压器为0.1，BLSTM为0.2。为了提高训练吞吐量，我们的批大小是动态确定的，这样我们就可以占用尽可能多的GPU内存。对于这项工作中的大多数实验，一批包含大约10000到20000帧，包括填充帧。我们使用32个NVIDIAP100GPU对模型进行最多100个时代的培训；培训通常在4天内完成。我们没有对变压器或BLSTM执行彻底的架构搜索。对于变压器，我们主要使用12层的变压器架构，其值为768：每个头的维度总是64，而FFN维度总是设置为4di。该模型的参数约为90M。对于BLSTM，我们遵循[24]并考虑两种架构，一种是每层每方向800个单元的5层BLSTM（约94M参数），另一种是1000个单元的6层BLSTM（约163M参数）[1]。*di公司*

训练变形金刚需要一些技巧。由于输入序列长度的计算量呈二次增长，我们将训练语句分割成不超过10秒的片段[2]。虽然这造成了训练和测试之间的不匹配，但初步结果表明，在较短的训练段上训练不仅提高了训练吞吐量，而且有助于最终的WER。我们还发现，变压器更容易过度拟合，因此需要一些正则化。我们发现SpecAugment[10]是有效的：没有SpecAugment，WER仅在3个时期后开始增加，而WER在SpecAugment的训练中继续提高。

使用Kaldi建立了一个完全优化的静态4-gram译码图。该译码图用于第一通译码和神经网络LM重构的n-最佳生成。测试集WER是使用基于开发集WER的最佳模型获得的[3]。在[38]之后，在相应的开发集[4]4.2位置嵌入的影响上分别选择test clean和test other的最佳检查点

在第一组实验中，我们研究了四种位置嵌入方法对基于变压器的声学模型的影响。在第一种方法中，我们每两帧进行一次叠加和跨步：它不会破坏变压器的排列不变性，因此表示为无。在第二种方法中，使用了原始变压器论文[14]中提出的正弦PE，它对绝对位置信息进行编码。在第三种方法，帧叠加中，我们将当前帧和接下来的8个未来帧叠加，然后进行一次Stripe2采样，形成一个新的变压器输入序列。注意，由于堆叠的帧与其相邻的堆叠帧部分重叠，因此排列不变性不再成立。该方法对相对位置信息进行编码。在第四种方法卷积中，我们在变压器层下面使用了两个VGG块[39]：每个VGG块包含两个连续的卷积层，每个卷积层带有一个3乘3的内核，后跟一个ReLu非线性和一个池层；第一个VGG块的卷积层使用32个通道，第二个VGG块增加到64个通道。Maxpooling是在一个2乘2的网格上执行的，跨步2在第一个块中，1在第二个块中。对于以10ms速率输入的80维特征向量序列，该VGG网络以20ms速率产生2560维特征向量序列。注意，VGG网络输出的每个特征向量的感知场由80ms左上下文和80ms右上下文组成，与帧叠加相同的右上下文长度。线性投影用于将特征向量投影到变压器接受的尺寸，在本例中为768。

表1：变压器位置嵌入（PE）的影响。

|  |  |  |
| --- | --- | --- |
| PE | 测试清洁 | 测试其他 |
| *没有* | 3.11 | 6.94 |
| *正弦波* | 3.13 | 6.67 |
| *帧堆叠* | 3.04 | 6.64 |
| *卷积* | 2.87 | 6.46 |

4.3变压器与BLSTM

在第二组实验中，我们比较了变压器结构和BLSTM。为了公平比较，我们尝试使用相似数量的参数建立基于变压器和BLSTM的模型。首先，我们将BLSTM模型BLSTM（800，5）与表1第3行中的transformer模型进行比较，BLSTM（800，5）是5层，每层每个方向有800个隐藏单元，因为它使用帧堆叠，所以称为Trf FS。为了能够比较基于变压器的最佳性能模型和卷积PE，我们将表1第4行中相同的VGG块与BLSTM结合，生成VGBLSTM（800，5）。最后，利用约163M的参数，建立了最大的vggBLSTM模型vggBLSTM（1000,6）。为了匹配该模型的参数个数，我们将变压器层数从12层增加到20层。如表2所示，基于transformerbased的模型在test clean和test other上的表现始终优于基于BLSTM的模型，分别为2-4%和7-11%。

4.4重复损失的影响

表2显示，只要将变压器的深度增加到20层，我们就可以获得大约5.5%的相对功耗降低（6.10 vs。

6.46). 受此启发，我们尝试增加transformer表2的数量：Librispeech基准上的架构比较

|  |  |  |  |
| --- | --- | --- | --- |
| 模型拱 | #参数（M） | 测试清洁 | 测试其他 |
| BLSTM（800,5） | 79 | 3.11 | 7.44 |
| Trf FS（768,12） | 91 | 3.04 | 6.64 |
| VGBLSTM（800,5） | 95 | 2.99 | 6.95 |
| vggTrf公司。(768,12) | 93 | 2.87 | 6.46 |
| VGBLSTM（1000,6） | 163 | 2.86 | 6.63 |
| vggTrf公司。(768, 20) | 149 | 2.77 | 6.10 |

进一步分层。为了使模型的大小易于管理，我们使用较小的嵌入维数512作为深变换器模型。我们最初的尝试并没有成功；深变形金刚模型（超过20层）经常被困在训练中，在很长一段时间内几乎没有进展。我们用[29]中使用的迭代损耗解决了这个问题：对第6/12/18个变压器层的输出嵌入进行非线性变换（投影到256维空间，然后进行线性变换，再进行相关非线性），并分别计算辅助CE损耗。这些额外的CE损失与原始CE损失（权重为0.3）进行插值。通过这种迭代损耗，我们能够训练一个24层变压器模型，在解码时只有81M的模型参数[5]，并且在vggTrf（768，12）基线上，在测试清洁和测试其他时分别获得7%和13%的功耗降低。

表3：使用迭代损耗训练深层变压器模型。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型拱 | Iter损耗 | 测试清洁 | 测试其他 |
| vggTrf公司。(768, 12) | 不 | 2.87              6.46 | |
| （参数：93M） | 是的 | 2.77              6.10 | |
| vggTrf公司。(512, 24) | 不 | 不收敛 | |
| （参数：81米） | 是的 | 2.66              5.64 | |

在这个vggTrf（512，24）模型的基础上，我们进一步进行了sMBR训练，在test-clean和test-other上略微提高到2.60%和5.59%。我们将我们的结果与一些已发表的Librispeech最新系统进行了比较，如表4所示：当使用标准4-gram LM进行解码时，我们的系统在测试清理和测试其他方面分别比以前最好的4-gram混合系统[24][6]降低了19%和26%。我们还在Librispeech基准提供的800M文本令牌上构建了一个类似于[16]中设置的transformer LM，并对第一次通过的解码输出执行了n-best重排序。据我们所知，我们的最终WER（2.26/4.85）是这个广泛使用的基准的最新结果。

表4:Librispeech与以前最佳结果的比较。“4g”表示使用4克LM；“NNLM”表示使用神经LM。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 拱门。 | 系统 | 流光溢彩 | 测试清洁 | 测试其他 |
| 拉斯维加斯 | Park等人[10]Karita等人[30] | NNLM+4g NNLM | 2.5 2.6 | 5.8 5.7 |
| 混合的 | 跑道[38] | 4克  +NNLM公司 | 3.8 2.3 | 8.8 5.0 |
| Han等人[41] | 4克  +NNLM公司 | 2.9 2.2 | 8.3  5.8 |
| Le等人[24] | 4克 | 3.2 | 7.6 |
| 我们的 | 4克  +NNLM公司 | 2.60 2.26 | 5.59  4.85 |

表5：强制转换器模型在推理过程中每层使用有限的右上下文（RC）。给定一个12层的变压器，10帧的RC转换为2.48秒的总前瞻。

|  |  |  |
| --- | --- | --- |
| 钢筋混凝土 | 测试清洁 | 测试其他 |
| ∞ | 2.87 | 6.45 |
| 50 | 3.01 | 7.12 |
| 20 | 3.29 | 8.10 |
| 10 | 3.65 | 9.01 |

4.5有限权利背景

到目前为止，所有基于变压器的实验都使用了完整的上下文。为了了解转换器在多大程度上依赖于未来的帧来导出当前帧的嵌入，我们采用vggTrf（768，12）模型（第4行，表2），并强制每一层在推理过程中关注一个固定的有限权利上下文。有趣的是，尽管这在训练和推理之间造成了很大的不匹配，但是如果正确的上下文帧的数目足够大，那么生成的系统仍然可以产生合理的wer。请注意，尽管每个层只需要有限的右上下文帧，但总的右上下文长度是由每个转换器层的右上下文长度相加的，因此我们最终仍然会有一个大的面向未来的展望窗口，这使得在流式ASR应用程序中使用的可能性降低。在未来的研究中，我们将研究基于变压器的具有流约束的声学模型。

4.6大规模实验

最后，我们进行了一个大规模的实验，我们的内部任务之一，英语视频ASR。该训练集由用户公开共享的137k小时视频（来自9416k个视频片段）组成；在我们的实验中只使用了这些视频的音频部分。这些数据是完全匿名的；转录者和研究者都不能访问任何用户可识别的信息。由于数据的性质，这是一个非常多样化和富有挑战性的任务。大约9小时（来自620个视频剪辑）的数据为dev set保留。3个测试集用于评估目的：一组8.5小时精心挑选的干净视频，一组19小时的干净视频和一组18.6小时的嘈杂视频。为了我们最初的评估目的，训练集和测试集被分成最多10秒的部分。

由于时间的限制，我们只构建了vggTrf（768，12）而没有迭代损失，并在此任务上构建了vggBLSTM（800，5）。表6显示，在这项任务中，提出的基于变压器的声学模型比VGBLSTM的性能好4.0-7.6%。我们将在今后的工作中汇报更多的成果。

表6：我们内部英语视频ASR任务的实验结果。

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 策划 | 清洁的 | 嘈杂的 |
| VGBLSTM（800,5） | 10.72 | 15.97 | 22.13 |
| vggTrf（768,12） | 9.90 | 15.26 | 21.25 |

# 5        讨论和结论

在这项工作中，我们提出并评估了基于变压器的混合语音识别声学模型。讨论并比较了几种模型选择。我们证明了transformer可以显著地优于BLSTM，并且在Librispeech基准上给出了最好的声学模型。对一个更大更具挑战性的数据集的初步研究也证实了我们的发现。

我们还有很多工作要做。例如，我们的实验并没有显示transformer的卓越性能在多大程度上来自于用自我关注代替重复，而transformer的其他建模技术也可以用来改进RNN[42]。相对于语音信号长度而言，二次增长的成本仍然是基于变压器的声学模型在实际中使用的主要障碍。这些问题将在我们今后的工作中加以研究。

# 6        工具书类

[1] G.Hinton，L.Deng，D.Yu，et al.，“语音识别中声学建模的深度神经网络”，IEEE信号处理杂志，第29卷，2012年。

[2] F.Seide、G.Li和D.Yu，“使用上下文相关深层神经网络的会话语音转录”，在Proc。Interspeech，2011年。

[3] H.Sak，A.Senior和F.Beaufays，“用于大规模声学建模的长-短期记忆递归神经网络结构”，过程。Interspeech，2014年。

[4] O.Abdel Hamid，A.Mohamed，H.Jiang等人，“用于语音识别的卷积神经网络”，IEEE/ACM音频、语音和语言处理学报，第22卷，第10期，第1533-15452014页。

[5] V.Peddinti、D.Povey和S.Khudanpur，“用于长时间上下文有效建模的时滞神经网络结构”，在Proc。Interspeech，2015年。

[6] 张国忠，江国华，魏国伟，戴立群，“无反馈的前向顺序记忆神经网络”，arXiv预印本附件十四：1510.02693, 2015.

[7] S.Hochreiter和J.Schmidhuber，“长-短期记忆”，《神经计算》，第9卷，第8期，第1735-17801997页。

[8] D.Bahdanau，J.Chorowski，D.Serdyuk等，“基于端到端注意的大词汇量语音识别”，过程。ICASSP，2016年。

[9] C.-C.Chiu，T.N.Sainath，Y.Wu，et al.，“具有序列到序列模型的最新语音识别”，过程。ICASSP，2018年。

[10] Park，W.Chan，Y.Zhang，et al.，“Specaugment:一种用于自动语音识别的简单数据增强方法”，arXiv预印本附件十四：1904.08779, 2019.

[11] Y.He，T.N.Sainath，R.Prabhavalkar等人，“移动设备的流式端到端语音识别”，过程。ICASSP，2019年。

[12] Y.Bengio，P.Simard等人，“学习梯度下降的长期依赖性是困难的”，IEEE神经网络学报，第5卷，第2期，第157-166页，1994年。

[13] R.Collobert、C.Puhrsch和G.Synnaeve，“Wav2letter:基于端到端convnet的语音识别系统”，arXiv预印本附件十四：1609.03193, 2016.

[14] A.Vaswani，N.Shazeer，N.Parmar等人，《关注是你所需要的一切》，过程。NIPS，2017年，第5998-6008页。

[15] J.Devlin，M.-W.Chang，K.Lee，et al.，“Bert：语言理解深度双向变压器的预培训”，arXiv预印本附件十四：1810.04805, 2018.

[16] A.Radford，K.Narasimhan，Tim S.等人，《通过生成性预训练提高语言理解》，2018年。

[17] 李东生，徐国荣，徐国荣，“语音转换器：语音识别的无重复序列到序列模型”，载于。ICASSP，2018年。

[18] M.Sperber，J.Niehues，G.Neubig等人，“自我注意声学模型”，arXiv预印本附件十四：1803.09519, 2018.

[19] 周国荣，董立军，徐国荣，徐国荣，“基于音节的汉语语音序列识别”，arXiv预印本附件十四：1804.10752, 2018.

[20] D.Povey，Hossein Hadian，P.Ghahremani，et al.，“asr的时间限制的自我注意层”，过程。ICASSP，2018年，第5874-5878页。

[21]J.Salazar、K.Kirchhoff和Z.Huang，“语音识别中连接主义时间分类的自我注意网络”，摘自Proc。ICASSP，2019年，第7115-7119页。

[22]H.A.Bourlard和N.Morgan，《连接主义语音识别：混合方法》，第247卷，Springer Science&Business Media。

[23]M.-Y.Hwang和X.Huang，“马尔可夫状态下的亚音阶建模”，在Proc。ICASSP，1992年，第1卷，第33-36页。

[24]D.Le，X.Zhang，W.Zheng，et al.，“从senones到chenones：用于混合语音识别的绑定上下文相关图形集”，arXiv预印本附件十四：1910.01493, 2019.

[25]J.Lei Ba、J.Kiros和G.E.Hinton，“层标准化”，arXiv预印本附件十四：1607.06450, 2016.

[26]D.Hendrycks和K.Gimpel，“高斯误差线性单位（gelus）”，arXiv预印本附件十四：1606.08415, 2016.

[27]A.Mohamed、D.Okhonko和L.Zettlemoyer，“asr卷积上下文变压器”，arXiv预印本附件十四：1904.11660, 2019.

[28]Y.Zhang、W.Chan和N.Jaitly，“用于端到端语音识别的非常深的卷积网络”，在Proc。ICASSP，2017年，第4845-4849页。

[29]A.Tjandra，C.Liu，F.Zhang，et al.，“似曾相识：深部变压器网络中的双重特征呈现和迭代损耗”，发表于ICASSP，2020年。

[30]S.Karita，N.Chen，T.Hayashi等人，“语音应用中变压器与RNN的比较研究”，arXiv预印本附件十四：1909.06317, 2019.

[31]R.Al Rfou，D.Choe，N.Constant等人，“具有更深层自我关注的字符级语言建模”，《AAAI人工智能会议记录》，2019年，第33卷，第3159-3166页。

[32]V.Panayotov，G.Chen，D.Povey，and S.Khudanpur，“Librispeech:基于公共领域有声图书的asr语料库”，in Proc。ICASSP，2015年，第5206-5210页。

[33]D.Povey，A.Ghoshal，G.Boulianne等人，“kaldi语音识别工具包”，IEEE自动语音识别与理解研讨会，2011年。

[34]T.Ko，V.Peddinti，D.Povey，et al.，“用于语音识别的音频增强”，in Proc。Interspeech，2015年。

[35]K.Vesely、A.Ghoshal、L.Burget和D.Povey，“深层神经网络的序列判别训练”，过程。Interspeech，2013年，第2013卷，第2345-2349页。

[36]O.Myle，E.Sergey，B.Alexei，F.Angela，et al.，“fairseq:序列建模的快速可扩展工具包”，摘自《NAACLHLT 2019年会议记录：演示》，2019年。

[37]D.P.Kingma和J.Ba，“Adam:随机优化方法”，arXiv预印本附件十四：1412.6980, 2014.

[38]C.Luscher，E.Beck，K.Irie，等人，“RWTH ASR系统用于Lib-¨riSpeech:混合vs注意-无数据增强”，arXiv预印本附件十四：1905.03072, 2019.

[39]K.Simonyan和A.Zisserman，“用于大规模图像识别的超深卷积网络”，arXiv预印本附件十四：1409.1556, 2014.

[40]Y.Zhang，G.Chen，D.Yu，et al.，“用于远程语音识别的高速公路长-短期记忆RNN”，过程。ICASSP公司。IEEE，2016年，第5755-5759页。

[41]K.J.Han，R.Prieto，K.Wu，and T.Ma，“使用扩展1d卷积的多流自我注意的最新语音识别”，arXiv预印本附件十四：1910.00716, 2019.

[42]M.Chen，O.Firat，A.Bapna，et al.，“两全其美：结合神经机器翻译的最新进展”，arXiv预印本附件十四：1804.09849, 2018.

[[1]](" \l "_ftnref1" \o ")通过将BLSTM中的参数数量增加到163M以上，我们没有得到进一步的改善。

[[2]](" \l "_ftnref2" \o ")这是通过使用现有的延迟控制BLSTM声学模型将音频与参考对齐来实现的。

[[3]](" \l "_ftnref3" \o ")我们还平均最后10个epoch检查点，以形成一个额外的候选。

[[4]](" \l "_ftnref4" \o ")这只是为了遵循前面的工作在[38]中设定的相同实验方案——两个测试集上的大多数实验结果，包括我们在表4中报告的最佳WER，实际上是通过相同的模型实现的。

[[5]](" \l "_ftnref5" \o ")有6个额外的参数只用于训练。

[[6]](" \l "_ftnref6" \o ")注意，[24]使用了LC-BLSTM[40]而不是全上下文BLSTM。