

Attention-Based Models for Speech Recognition

基于注意的语音识别模型



Jan Chorowski, University of Wrocław, Poland

Dzmitry Bahdanau, Jacobs University Bremen, Germany

Dmitriy Serdyuk, Universit´e de Montr´eal

Kyunghyun Cho, Universit´e de Montr´eal

Yoshua Bengio, Universit´e de Montr´eal, CIFAR Senior Fellow

摘要

Recurrent sequence generators conditioned on input data through an attention mechanism have recently shown very good performance on a range of tasks including machine translation, handwriting synthesis [1, 2] and image caption generation [3]. We extend the attention-mechanism with features needed for speech recognition. We show that while an adaptation of the model used for machine translation in [2] reaches a competitive 18.7% phoneme error rate (PER) on the TIMIT phoneme recognition task, it can only be applied to utterances which are roughly as long as the ones it was trained on. We offer a qualitative explanation of this failure and propose a novel and generic method of adding location-awareness to the attention mechanism to alleviate this issue. The new method yields a model that is robust to long inputs and achieves 18% PER in single utterances and 20% in 10-times longer (repeated) utterances. Finally, we propose a change to the attention mechanism that prevents it from concentrating too much on single frames, which further reduces PER to 17.6% level.

通过注意机制以输入数据为条件的循环序列生成器最近在机器翻译、手写合成[1，2]和图像字幕生成[3]等一系列任务中表现出非常好的性能。我们用语音识别所需的特征来扩展注意机制。我们发现，尽管[2]中用于机器翻译的模型在TIMIT音位识别任务中达到了18.7%的音位错误率（PER），但它只能应用于与训练时长度大致相同的话语。我们提供了一个定性的解释，并提出了一种新的和通用的方法，增加位置意识的注意机制，以缓解这一问题。新方法产生了一个对长输入具有鲁棒性的模型，在单句话中达到18%，在10倍长（重复）的话中达到20%。最后，我们提出了一个改变注意机制的方法，以防止注意力过于集中在单个帧上，从而进一步将PER降低到17.6%。

2015.06.24 arxiv

# 1        介绍

近年来，基于注意的递归网络已成功地应用于各种各样的任务，如手写合成[1]、机器翻译[2]、图像字幕生成[3]和视觉对象分类[4]，这些模型通过在每一步选择相关内容来迭代处理输入。这一基本思想极大地扩展了端到端训练方法的适用范围，例如，使构建具有外部存储器的网络成为可能[6，7]。[[1]](" \l "_ftn1" \o ")

我们介绍了基于注意的循环网络的扩展，使其适用于语音识别。学习识别语音可视为学习生成给定另一序列（语音）的序列（转录）。从这个角度来看，它类似于机器翻译和手写合成任务，基于注意的方法被认为是合适的[2，1]。然而，与机器翻译相比，语音识别的主要区别在于需要更长的输入序列（数千帧而不是几十个单词），这就给区分单个话语中相似的语音片段带来了挑战。它也不同于手写合成，因为输入序列噪声大，结构不清晰。基于这些原因，语音识别是一个有趣的测试平台，用于开发能够处理长时间和噪声输入的新的基于注意的架构。[[2]](" \l "_ftn2" \o ")

将基于注意的模型应用于语音识别，也是构建完全端到端可训练语音识别系统的重要一步，这是一个活跃的研究领域。主要的方法仍然是基于混合系统，包括深层神经声学模型、三音HMM模型和n-gram语言模型[8,9]。这需要手工制作的发音词典和音素词典，以及一个多阶段的训练过程，使组件协同工作。最近报道了一种由CTC训练的神经网络和语言模型组成的无隐马尔可夫模型识别器的优秀结果[10]。然而，语言模型只是在这项工作的最后一个阶段才被添加的，因此留下了一个悬而未决的问题，即在训练过程中，一个声学模型能从意识到一个语言模型中受益多少。

在这篇论文中，我们使用广泛使用的TIMIT数据集评估基于注意的音素识别模型。在产生输出序列（音素）的每个时间步，注意机制选择或加权由训练的特征提取机制在输入序列（语音帧）中潜在的所有时间步产生的信号。然后，加权特征向量有助于调节输出序列的下一个元素的生成。由于该数据集中的话语非常短（大多在5秒以下），因此我们衡量了所考虑的模型在识别通过人工连接现有话语而产生的更长话语的能力。

我们从[2]中提出的机器翻译任务模型开始，作为基线。这个模型似乎完全容易受到类似语音片段问题的影响，但尽管我们期望它在原始测试集上具有竞争力，达到了18.7%的音素错误率（PER）。然而，它的性能下降很快与较长，连接发言。我们提供的证据表明，该模型适用于跟踪其识别的内容在输入序列中的绝对位置，这种策略对于来自原始测试集的短句是可行的，但本质上是不可伸缩的。

为了避免这种不期望的行为，在本文中，我们建议修改注意机制，使其明确考虑（a）上一步的焦点位置，如[6]所示，以及（b）输入序列的特征，如[2]所示。这是通过添加辅助卷积特征作为注意机制的输入来实现的，辅助卷积特征是通过使用可训练滤波器卷积前一步的注意权重来提取的。我们发现，具有这种卷积特征的模型在所考虑的任务上表现得更好（18.0%PER）。更重要的是，具有卷积特征的模型能够很好地识别出比训练集长很多倍的话语，每次识别率始终低于20%。

因此，这项工作的贡献是三倍的。首先，我们提出了一种新的基于注意机制的纯神经语音识别结构，其性能与TIMIT数据集上的传统方法相当。此外，我们还提出了一种在注意机制中加入位置感知的通用方法。最后，我们对注意机制进行了改进，避免了将注意力集中在单个帧上，从而避免了获得不太“有效的训练实例”，使PER下降到17.6%。

# 2        基于注意的语音识别模型

2.1总体框架

基于注意的递归序列发生器（ARSG）是一种递归神经网络，它从输入随机生成一个输出序列（y1，…，yT）。在实践中，通常由编码器处理，编码器输出更适合注意力机制使用的顺序输入表示=（h1，…，hL）。*十十小时*

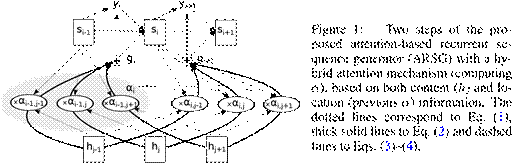
在本文中，输出是一个音素序列，输入=（x1，…，xL0）是一个特征向量序列。每个特征向量都是从音频帧的一个小重叠窗口中提取出来的。编码器被实现为一个深度双向递归网络（BiRNN），以形成长度=L0的序列表示。*是的十小时我*

在第-步，ARSG通过关注以下相关元素来生成输出：*我彝语小时*

*αi*=参加（si−1，αi−1，h）(1)

*我*

|  |  |
| --- | --- |
| *克我*=Xα*i、 日本小时日本*  *日本*=1 | (2) |
| *彝语*∼生成（si−1，gi）， | (3) |



其中是递归神经网络的第（i−1）个状态，我们称之为生成器，∈Ris是注意权重的向量，也通常称为对齐[2]。使用[4]中的术语，我们称之为一瞥。该步骤通过计算新的发电机状态完成：*硅*−1*αi我胃肠道*

*硅*=复发（si−1，g，yii）(4)

长-短期记忆单位（LSTM，[11]）和门控复发单位（GRU，[12]）通常被用作复发激活，我们称之为复发。该过程在图1中以图形方式示出。

受[6]的启发，我们区分了基于位置、基于内容和混合注意机制。式（1）中描述了最普遍的混合注意。如果该术语从参数中删除，即，=Attend（si−1，h），我们称之为基于内容的（参见，例如，[2]或[3]）。在这种情况下，通常通过对每个元素分别打分并对分数进行标准化来实现：*出席αi*−1*出席αi出席小时*

|  |  |
| --- | --- |
| *电子i、 日本*=分数（s−1），*我，小时日本* | (5) |
| *.* | (6) |

*日本*=1

这种方案的主要局限性在于，相同或非常相似的元素不论在序列中的位置如何，都会得到相等的分数。这是上文提出的“类似的讲话片段”问题。通常，通过编码器（例如BiRNN[2]或深度卷积网络[3]）来部分地缓解该问题，该编码器将上下文信息编码到每个元素中。然而，元素的容量总是有限的，因此上下文消歧只能在有限的范围内实现。*小时小时小时*

或者，基于位置的注意机制从发生器状态和先前的对准计算对准，仅使得=参加（si−1，αi−1）。例如，Graves[1]在他的手写合成模型中使用了基于位置的注意机制，使用了高斯混合模型。在语音识别的情况下，这种基于位置的注意机制只能使用单个音素来预测后续音素之间的距离，我们认为这是很困难的，因为这个数量的变化很大。*αi硅*−1

对于基于内容和基于位置机制的局限性，我们认为混合注意机制是语音识别的一种自然选择。非正式地说，我们想要一个注意力模型，它使用前面的对齐来选择一个简短的元素列表，从中基于内容的注意力，在Eqs中。（5） –（6），将选择相关的，而不会混淆。*αi*−1*小时*

2.2建议模型：具有卷积特征的ARSG

我们从基于ARSG的模型出发，结合文献[2]提出的基于内容的注意机制。这个模型可以用等式来描述。（5） –（6），其中

*艾未未，j*=w>tanh（Wsi−1+V hj+b）。(7)

*w型*是向量，是矩阵。*b类W型五*

我们将原始模型中基于内容的注意机制扩展到位置感知，使其考虑到前一步产生的对齐。首先，我们通过用矩阵R×R卷积，为先前对齐的每个位置提取向量∈rf：*k公司fi、jk公司日本αi*−1*F级*∈*k公司*

*金融机构*=F\*αi−1。(8)

这些额外的向量然后由评分机制使用：*fi、j艾未未，j*

*艾未未，j*=w>tanh（Wsi−1+V hj+Ufi，j+b）(9)

2.3分数归一化：锐化和平滑

公式（6）中的标准化有三个潜在问题。

首先，当输入序列较长时，由于归一化分数均为正且和为1，一瞥可能包含来自许多无关特征向量的噪声信息。这使得拟议的ARSG很难每次都清楚地集中在几个相关的帧上。第二，注意机制需要在每次解码单个输出时考虑所有帧，同时解码长度的输出，导致计算复杂度（LT）。当输入语句很长时，这可能很容易变得非常昂贵（对于机器翻译来说，这个问题不太严重，因为在这种情况下，输入序列是由单词组成的，而不是20ms的声学帧）。*小时胃肠道hj公司αi，j我我彝语T型O*

另一方面，公式（6）中使用的softmax标准化倾向于只关注单个特征向量。这会阻止模型聚合多个顶部得分的帧以形成一瞥。*hj公司胃肠道*

锐化通过“锐化”分数，有一种直截了当的方法来解决第一个问题。锐化权重的一种方法是在softmax函数中引入逆温度1，以便*αi，jβ >*

, *我*

*一i、 日本*=exp（βe）Xexp（βe），*i、 日本i、 日本*

*日本*=1

或者根据分数只保留最上面的帧并重新标准化。然而，这些锐化方法仍然需要我们每次计算每一帧的分数（（LT）），并且它们恶化了第二个问题，即焦距过窄。*k公司O*

我们还提出并研究了一种窗口技术。每次，注意机制只考虑整个序列的一个子序列=（hp）*我小时*˜*我*−宽*，…，马力我*+w−1级*小时*，其中是预定义的窗口宽度和*圆周率*是路线的中线。没有计算∈/hare的分数，导致较低的复杂性（L+T）。这种加窗技术类似于使用顶部帧，同样具有锐化效果。*αi*−1*hj公司*˜*Ok公司*

提出的基于窗口化的锐化方法可以用于训练和评估。后来，在实验中，我们只考虑在评估时使用它的情况。

平滑我们观察到，建议的锐化方法确实有助于长话语。然而，所有这些因素，尤其是选择得分最高的框架，都会对模型在以短句为主的标准发展集上的表现产生负面影响。这一观察结果让我们假设，它有助于该模型从多个得分最高的帧聚合选择。从某种意义上说，当考虑更多的输入位置时，这将为模型的输出部分带来更多的多样性，即更有效的训练示例。为了促进这一效果，我们将等式（6）中softmax函数的无界指数函数替换为有界logistic sigmoid，以便*σ*

, *我一i、 日本*=σ（e）*i、 日本*)Xσ（e）*i、 日本*) .

*日本*=1

这有平滑注意力机制发现的焦点的效果。

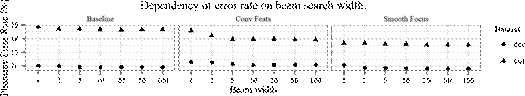


图2：解码性能w.r.t.波束大小。为了进行严格的比较，如果解码不能产生heosi，我们认为它是错误的，没有用较大的波束大小重试。这些模型，特别是具有平滑聚焦的模型，即使在束宽只有1的情况下也表现良好。

# 3        相关工作

基于连接论时间分类（CTC，[13]）及其扩展RNN变换器[14]的语音识别器最接近本文考虑的ARSG模型。他们遵循了早期关于在序列上进行端到端可训练的深度学习的工作，梯度信号流经对齐过程[15]。它们在音素识别任务中表现良好[16]。此外，最近发现CTC能够直接从语音中转录文本，而无需任何中间语音表示[17]。

所考虑的ARSG在两个方面与CTC和RNN传感器不同。首先，尽管注意机制决定了输入和输出序列的对齐，但是CTC和RNN传感器将对齐视为潜在的随机变量，在该变量上执行MAP（最大后验概率）推理。ARSG对准机制的这种确定性使得波束搜索过程更加简单。此外，我们根据经验观察到，小得多的波束宽度可用于确定性机制，这允许更快的解码（见第。4.2和图2）。第二，CTC和RNN传感器的对准机制都被限制为“单调”的，以保持对准的边缘化易于处理。另一方面，所提出的注意机制可以导致非单调对齐，这使得它适用于除语音识别以外的更广泛的任务。

文[6]还提出了一种基于卷积运算的神经图灵机混合注意模型。在每个时间步，NTM计算基于内容的注意权重，然后将其与预测的移位分布卷积。与NTM的方法不同，这里提出的混合机制让学习知道基于内容和基于位置的寻址是如何通过一个深度的参数函数组合的（见等式（9））

Sukhbatar等人[18]描述了一种类似的混合注意机制，其中位置嵌入被用作注意模型的输入。这种方法有一个重要的缺点，即模型不能处理比训练时更长的输入序列。另一方面，我们的方法可以很好地处理比训练时长很多倍的序列。5.)

# 4        实验装置

我们严格遵循了[16]中的程序。所有实验都是在TIMIT语料库上进行的[19]。我们使用了Kaldi[20]TIMIT s5配方中的train dev测试。我们在标准的462说话人集上进行了训练，去掉了所有SA语句，并使用了50说话人的dev集来提前停止。我们测试了24扬声器核心测试集。所有的网络都被训练在40个mel尺度滤波器组特征上，加上每帧的能量，以及第一和第二时间差，每帧总共产生123个特征。每个特征在训练集中被重新调整为零均值和单位方差。网络是在61个电话上训练的，每个目标序列都附加了一个额外的“序列结束”标记。类似地，我们在每个输入序列的末尾附加一个全零帧来表示话语的结束。解码是使用61+1音素集进行的，而评分是在39音素集上进行的。

4.1培训程序

ARSG模型的一个特点是不同的参数子集被重复使用的次数不同，编码器的重复使用次数不同，注意权重的重复使用次数也不同*我书信电报T型*

迈克尔用蜡笔给卧室的墙壁上色。

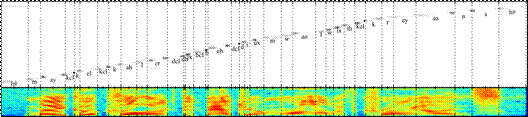


图3：基线模型生成的路线。垂直条表示TIMIT的地面电话位置。上面图像的每一行指示由注意机制选择的帧以发射电话符号。网络已经清楚地学会了从左到右排列，倾向于稍微向前看，并且不会混淆重复的“kclk”短语。最好看彩色。

ARSG的参数。这使得导数w.r.t.参数的尺度变化很大，我们使用自适应学习率算法AdaDelta[21]来处理它，AdaDelta[21]有两个超参数和。所有权重矩阵均由正态高斯分布初始化，其标准差设置为0.01。回归权进一步正交化。*ρ*

由于TIMIT是一个相对较小的数据集，适当的正则化是至关重要的。我们使用自适应权重噪声作为主要的正则化器[22]。我们首先用最大范数1的列范数约束[23]训练我们的模型，直到得到最低的发展负对数似然完成。期间这一次，和分别设置为10−8和0.95。在这一点上，我们开始使用自适应权重噪声，并降低了模型的复杂度成本[[3]](" \l "_ftn3" \o ") *ρ信用证*乘以10，同时禁用列规范约束。一旦达到新的最低开发对数可能性，我们就用较小的概率对模型进行微调，直到我们没有观察到100K权重更新的开发音素错误率（PER）的改善。在整个培训过程中使用了批量1。

4.2评估模型详情

我们用不同的注意机制来评估arsg。编码器是一个三层BiRNN，每个方向上有256个GRU单元，并使用512个顶层单元的激活作为表示。发电机有一个256个GRU单元的循环层。在式（3）中，有64个最大输出单位的隐藏层。将编码器和发生器的初始状态作为附加参数处理。*小时生成*

我们的基线模型是一个纯粹基于内容的注意机制（见等式）。（5） –（7）式（7）中的评分网络有512个隐藏单位。其他两个模型使用等式（8）中的卷积特征，其中=10和=201。其中一个使用秒平滑。2.3.*k公司右*

解码过程在解码过程中使用了从左到右的音素序列波束搜索[24]。发射“序列结束”标记heosi时，波束搜索停止。我们从波束宽度10开始，当网络无法产生波束更窄的heosi时，将其增加到40。如图2所示，用较宽波束进行解码几乎没有益处。

# 5        结果

所有模型均达到了竞争性PER（见表1）。使用卷积特征，我们可以看到相对于基线的相对改善3.7%，使用平滑的进一步改善5.9%。

令我们惊讶的是（见第。2.1.），基线模型学会正确对齐。图3显示了基线模型在具有重复音素的序列（话语FDHC0 SX209）上产生的对齐，这表明基线模型不会被短距离重复混淆。我们还可以从图中看出，它更喜欢选择靠近开头的帧或表1：音素错误率（PER）。粗体PER对应于基于注意的递归序列发生器（ARSG）的最佳错误率，ARSG结合了卷积注意特征和平滑聚焦。



|  |  |  |
| --- | --- | --- |
| 基线模型 | 15.9% | 18.7% |
| 基线+转换特性 | 16.1% | 18.0% |
| 基线+变换功能+平滑聚焦 | 15.8% | 17.6% |
| RNN传感器[16] | 不适用 | 17.7% |

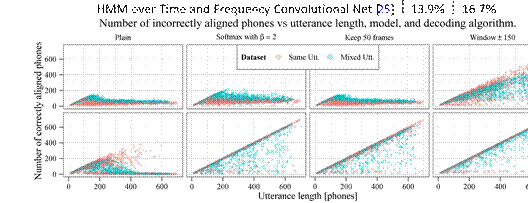


图4：连接话语的强制对齐结果。每个点代表一个单独的话语，它是由同一个话语或随机选择的不同话语的多个副本串联而成的。我们清楚地看到，当混合注意机制与建议的锐化技术相结合时，达到了最高的鲁棒性（见右下图）

甚至在作为数据集的一部分提供的音素位置之前。其他模型产生的对齐在视觉上非常相似。

5.1长话语的强制对齐

基线模型的良好性能使我们面临这样一个问题：它如何区分相似音素序列的重复，以及它如何可靠地解码重复次数较多的较长序列。我们创建了两个长话语的数据集；一个通过重复每个测试话语，另一个通过连接随机选择的话语。在这两种情况下，波形交叉褪色，插入0.05秒静音作为“pau”电话。我们连接了多达15个句子。

首先，我们通过强制生成器发出正确的音素来检查这些较长的语句的强制对齐。如果90%的对齐权重位于每侧延伸20帧的地面真音素窗口内，则每个对齐都被认为是正确的。在这个定义下，除了图3所示的heosi之外的所有电话都正确地对齐。

图4的第一列示出了正确对齐的帧的数目w.r.t.对于一些考虑的模型的话语长度（以帧为单位）。我们可以看到，当一个单独的话语被重复时，基线模型能够解码多达120个电话的序列，当不同的话语被串联时，能够解码多达150个电话的序列。即使失败了，它也正确地对准了大约50部手机。另一方面，具有卷积特征的混合注意机制的模型能够对齐长达200个电话的序列。然而，一旦它开始出现故障，该型号就无法校准几乎所有的手机。具有平滑特征的模型与仅具有卷积特征的模型表现相似。

我们检查了失败的排列，以了解这两种不同的失败模式。一些例子见补充材料。

我们发现，基线模型正确地排列了大约40首手机，然后跳转到录音结束，并在最后10首手机上循环。这种行为表明它学会了跟踪它在源序列中的大致位置。但是，跟踪能力仅限于训练期间观察到的长度。一旦跟踪器饱和，它就会跳到录音的结尾。

                                                基于长语音解码算法的音素误码率研究

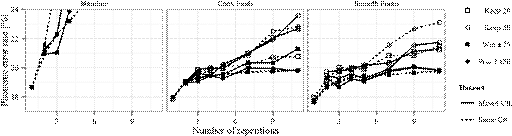


图5：解码长序列时获得的音素错误率。每一个网络被解码与对齐锐化技术，产生适当的强制对齐。提出的ARSG对话语长度的鲁棒性明显高于基准值。

相反，当位置感知网络出现故障时，它只是停止对齐-没有为每部手机选择特定的帧。我们将此行为归因于第节中讨论的噪声瞥见问题。2.3. 对于长话语，有许多不相关的框架会对分配给正确框架的权重产生负面影响。根据这一推测，位置感知网络在重复相同的话语（所有帧在某种程度上是相关的）上比在串联不同的话语（每个未对齐的帧都是不相关的）上工作得稍微好一些。

为了获得更多的洞察力，我们应用了第节中描述的对齐锐化方案。2.3. 在图4的其余列中，我们看到锐化方法有助于位置感知网络找到正确的对齐，而它们对基线网络几乎没有影响。窗口技术有助于基线和位置感知网络，位置感知网络可以正确对齐几乎所有序列。

在视觉检查中，我们注意到在很长的话语中间，基线模型被窗口内类似内容的重复所混淆，并且这种混淆在开始时没有发生。这支持了我们上面的推测。

5.2解码长话语

我们评估了长序列上的模型。每一个模型被解码使用对齐锐化技术，这有助于获得适当的强制对齐。结果如图5所示。基线模型无法解码长话语，即使使用窄窗口来约束它产生的对齐。另外两个位置感知网络能够解码由多达11个测试语句串联而成的语句。当注意机制在每一步都看到整个输入序列时，可能是因为它更像训练条件，所以在更宽的窗口中可以获得更好的结果。在宽窗口条件下，两个网络在长话语上的得分都在20%左右，这表明所提出的位置感知注意机制可以扩展到比训练集中的序列长得多的序列，只需要在解码阶段进行微小的修改。

# 6        结论

我们提出并评估了一种基于混合注意机制的端到端可训练语音识别架构，该架构结合了内容和位置信息，以便在输入序列中选择下一个位置进行解码。该模型的一个理想特性是，它能识别比训练时长得多的话语。在未来，我们希望该模型能够直接从语音中识别文本[10,17]，在这种情况下，将单语语言模型纳入ARSG体系结构可能变得非常重要[26]。

这项工作为注意机制提供了两个新的思路：一个更好的标准化方法产生更平滑的对齐和一个从先前对齐中提取和使用特征的通用原则。这两种方法都有可能应用于语音识别以外的领域。例如，所提出的注意可以在神经图灵机中使用而无需修改，或者通过使用二维卷积而不是一维卷积来改进图像字幕生成[3]。

致谢

所有实验均使用Theano[27,28]、PyLearn2[29]和Blocks[30]库进行。

作者感谢以下机构对研究资助和计算支持的支持：国家科学中心（波兰）、NSERC、魁北克省加尔各答、加拿大计算机协会、加拿大研究主席和CIFAR。Bahdanau还感谢Planet Intelligent Systems GmbH和Yandex。

# 工具书类

[1] 亚历克斯·格雷夫斯。用递归神经网络生成序列。附件十四：1308.0850，2013年8月。

[2] Dzmitry Bahdanau、Kyunghyun Cho和Yoshua Bengio。联合学习对齐和翻译的神经机器翻译。附件十四：1409.0473，2014年9月。

[3] Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhutdinov、Richard Zemel和Yoshua Bengio。展示、参与和讲述：视觉注意力的神经图像字幕生成。附件十四：1502.03044，2015年2月。

[4] 沃洛德迈尔·姆尼赫，尼古拉斯·希斯，亚历克斯·格雷夫斯，等。视觉注意的重现模型。《神经信息处理系统的进展》，第2204–2212页，2014年。

[5] 简·乔洛夫斯基、德米特里·巴达瑙、赵京贤和本吉奥。使用基于注意的递归神经网络的端到端连续语音识别：第一个结果。附件十四：1412.1602[cs，stat]，2014年12月。

[6] 亚历克斯·格雷夫斯、格雷格·韦恩和伊沃·丹尼尔卡。神经图灵机器。附件十四：1410.5401, 2014.

[7] 杰森·韦斯顿、苏米特·乔普拉和安托万·博尔德斯。记忆网络。附件十四：1410.3916, 2014.

[8] 马克·盖尔斯和史蒂夫·杨。隐马尔可夫模型在语音识别中的应用。找到了。趋势信号过程，1（3）：195–304，2007年1月。

[9] G.Hinton、Li Deng、Dong Yu、G.E.Dahl、Mohamed、N.Jaitly、Senior、V.Vanhoucke、P.Nguyen、T.N.Sainath和B.Kingsbury。语音识别声学建模的深层神经网络：四个研究小组的共同观点。IEEE信号处理杂志，29（6）：82–972012年11月。

[10] Awni Hannun、Carl Case、Jared Casper、Bryan Catanzaro、Greg Diamos、Erich Elsen、Ryan Prenger、Sanjeev Satheesh、Shubho Sengupta、Adam Coates等人，《Deepspeech:扩展端到端语音识别》。arXiv预印本附件十四：1412.5567, 2014.

[11] S.Hochreiter和J.Schmidhuber。长期短期记忆。神经质的。《计算机》，9（8）：1735–17801997年。

[12] 赵京铉、巴特·范·梅里恩波尔、卡格拉·古尔切赫、费提·布加勒斯、霍尔格·施文克和约书亚·本吉奥。统计机器翻译中使用RNN编码解码器学习短语表示。在EMNLP 2014中，2014年10月。出现。

[13] 亚历克斯·格雷夫斯、圣地亚哥·费尔南德斯、福斯蒂诺·戈麦斯和于尔根·施密杜伯。连接主义时间分类：用递归神经网络标记未分段的序列数据。在ICML-062006年。

[14] 亚历克斯·格雷夫斯。用递归神经网络进行序列转导。在ICML-12，2012年。

[15] Y.LeCun、L.Bottou、Y.Bengio和P.Haffner。梯度学习在文档识别中的应用。过程。IEEE，1998年。

[16] 亚历克斯·格雷夫斯、阿卜杜勒·拉赫曼·穆罕默德和杰弗里·辛顿。基于深度递归神经网络的语音识别。在ICASSP 2013中，第6645–6649页。IEEE，2013年。

[17] 亚历克斯·格雷夫斯和纳夫迪普·贾特利。面向端到端的递归神经网络语音识别。ICML-14，第1764–1772页，2014年。

[18] Sainbayar Sukhbatar、Arthur Szlam、Jason Weston和Rob Fergus。弱监督记忆网络。arXiv预印本附件十四：1503.08895, 2015.

[19] J.S.Garofolo、L.F.Lamel、W.M.Fisher、J.G.Fiscus、D.S.Pallett和N.L.Dahlgren。DARPA TIMIT声学-语音连续语音语料库，1993年。

[20] Daniel Povey、Arnab Ghoshal、Gilles Boulianne、Lukas Burget、Ondrej Glembek、Nagendra Goel、Mirko Hannemann、Petr Motlicek、Yanmin Qian、Petr Schwarz等。kaldi语音识别工具包。进行中。ASRU，第1-4页，2011年。

[21]马修·D·泽勒。ADADELTA：一种自适应学习率方法。附件十四：1212.5701, 2012.

[22]亚历克斯·格雷夫斯。神经网络的实用变分推理。J.Shawe Taylor、R.S.Zemel、P.L.Bartlett、F.Pereira和K.Q.Weinberger编著，神经信息处理系统进展24，第2348-2356页。Curran Associates，Inc.，2011年。

[23]杰弗里·辛顿（Geoffrey E Hinton）、尼蒂斯·斯利瓦斯塔瓦（Nitish Srivastava）、亚历克斯·克里兹耶夫斯基（Alex Krizhevsky）、伊利亚·萨茨基弗（Ilya Sutskever）和罗斯兰·萨拉胡季诺夫。通过防止特征检测器的共适应来改进神经网络。arXiv预印本附件十四：1207.0580, 2012.

[24]伊利亚·萨茨凯、奥里奥尔·维尼亚和库克诉勒。用神经网络进行序列间学习。arXiv预印本附件十四：1409.3215, 2014.

[25]拉兹洛托斯。基于卷积神经网络的电话识别中时域卷积和频域卷积的结合。在ICASSP 2014中，第190–194页，2014年。

[26]Caglar Gulcehre、Orhan Firat、Kelvin Xu、Kyunghyun Cho、Loic Barrault、Huei Chi Lin、Fethi Bogares、Holger Schwenk和Yoshua Bengio。单语语料库在神经机器翻译中的应用。arXiv预印本附件十四：1503.03535, 2015.

[27]詹姆斯·柏格斯特拉、奥利维尔·布鲁勒克斯、弗雷德·埃里克·巴斯蒂安、帕斯卡·兰布林、拉兹万·帕斯卡努、纪尧姆·德斯贾丁斯、约瑟夫·图里安、大卫·瓦尔德·法利和约书亚·本吉奥。Theano:CPU和GPU数学表达式编译器。在Python科学计算会议（SciPy）上，2010年6月。口头陈述。

[28]弗雷德•埃里克•巴斯蒂安、帕斯卡•兰布林、拉兹万•帕斯卡努、詹姆斯•伯格斯特拉、伊恩•J•古德费罗、阿诺•伯杰伦、尼古拉斯•布查德和约书亚•本吉奥。新功能和速度改进。深度学习和无监督特征学习NIPS 2012研讨会，2012年。

[29]Ian J.Goodfello、David Warde Farley、Pascal Lamblin、Vincent Dumoulin、Mehdi Mirza、Razvan Pascanu、James Bergstra、Fred´eric Bastien和Yoshua Bengio。´Pylearn2：机器学习研究图书馆。arXiv预印本附件十四：1308.4214, 2013.

[30]巴特·范·梅里恩波尔（Bart van Merrienboer）、德米特里·巴达瑙（Dzmitry Bahdanau）、文森特·杜穆林（Vincent Dumoulin）、德米特里·塞尔杜克（Dmitry Serdyuk）、大卫·瓦尔德·法利（David Warde Farley）、扬·乔洛夫斯基（Jan–Choro。积木与燃料：深度学习的框架。附件十四：1506.00619[cs，stat]，2015年6月。

# A附加数字

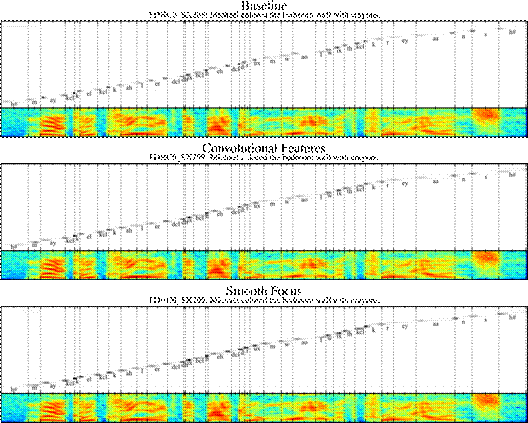


图6:FDHC0 SX209测试话语的评估模型产生的对齐。垂直条表示TIMIT的地面电话位置。上面图像的每一行指示由注意机制选择的帧以发射电话符号。与图3比较。在正文中。

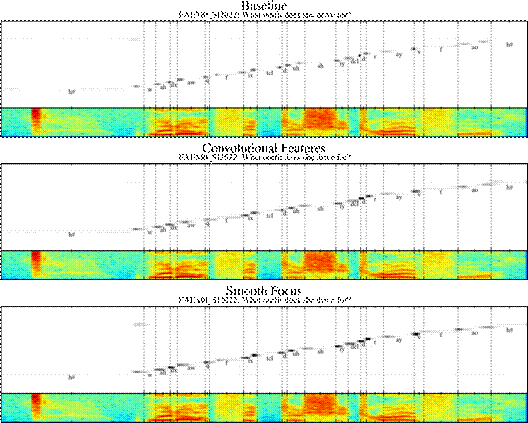


图7:FAEM0 SI2022列车语音评估模型产生的对齐。垂直条表示TIMIT的地面电话位置。上面图像的每一行指示由注意机制选择的帧以发射电话符号。与图3比较。在正文中。

错误排列的电话数量与话语长度和型号。

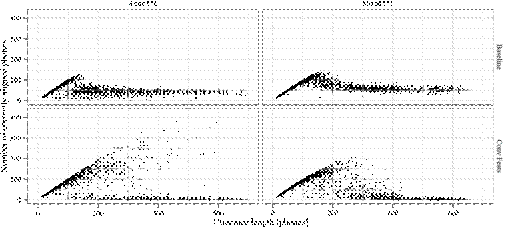
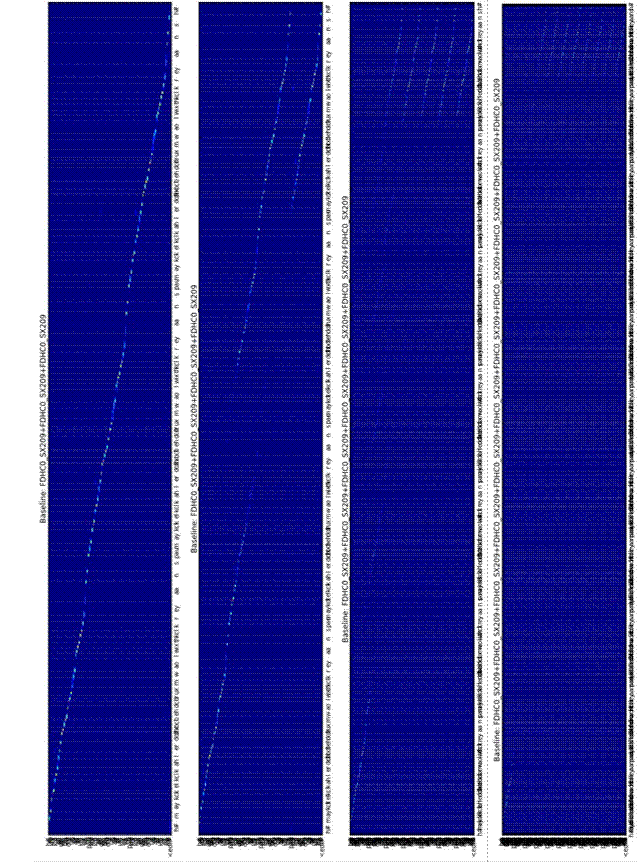
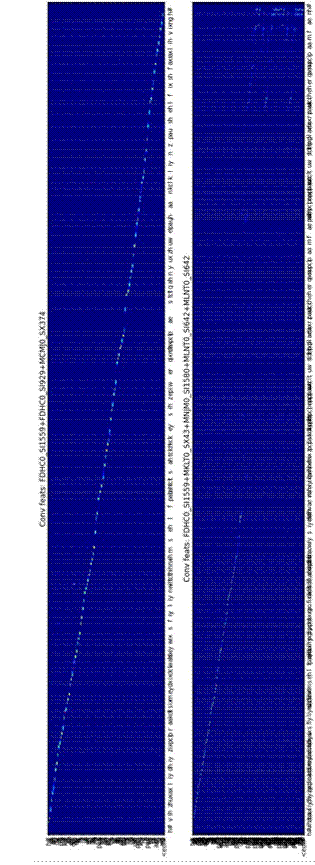
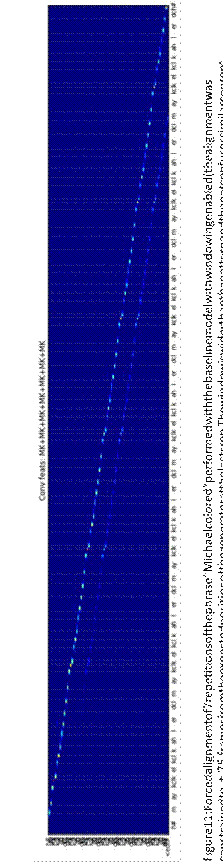
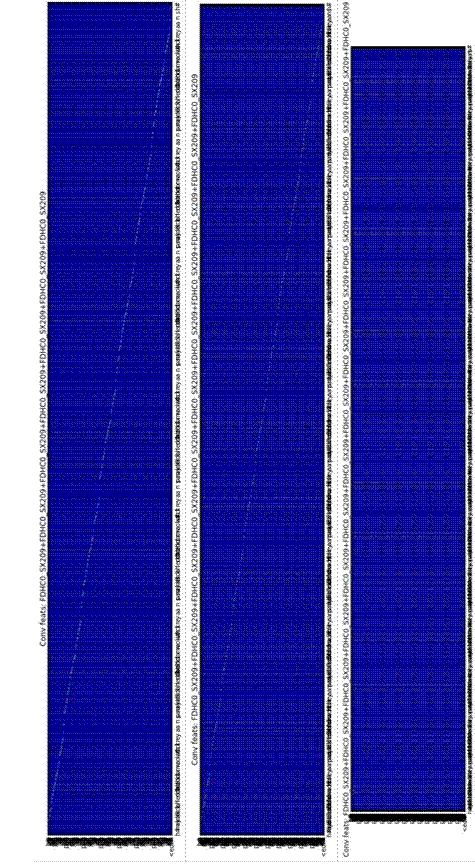


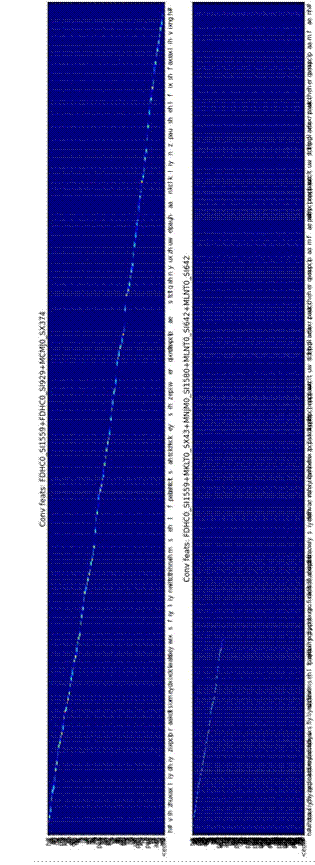
图8:ARSG两种故障模式的特写。力对齐连接的TIMIT语句的结果。每个点代表一个单独的话语。左侧面板显示了相同语句的串联结果。右边的面板显示了随机选择的话语串联的结果。我们比较了具有基于内容的注意机制（顶行）和使用卷积特征的混合注意机制（底行）的基线网络。虽然这两种模型都不能正确地对齐长序列，但它们以不同的方式失败：基线网络总是对齐大约50个电话，而位置感知网络无法对齐任何电话。与图4比较形成本文的主体。











# B实验的详细结果

表2：各种修改解码时的音素错误率。与本文的图5进行比较。

                                            普通保持1保持10保持50=2赢。±75胜。±150*β*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 基线 | 开发测试 | 15.9%  18.7% | 17.6%  20.2% | 15.9%  18.7% | 15.9%  18.7% | 16.1%  18.9% | 15.9%  18.7% | 15.9%  18.6% |
| 转换专长 | 开发测试 | 16.1%  18.0% | 19.4%  22.3% | 16.2%  17.9% | 16.1%  18.0% | 16.7%  18.7% | 16.0%  18.0% | 16.1%  18.0% |
| 平滑聚焦 | 开发测试 | 15.8%  17.6% | 21.6%  24.7% | 16.5%  18.7% | 16.1%  17.8% | 16.2%  18.4% | 16.2%  17.7% | 16.0%  17.6% |

[[1]](" \l "_ftnref1" \o ")这项工作的早期版本在NIPS 2014深度学习研讨会上发表[5]。

[[2]](" \l "_ftnref2" \o ")在第节中有更详细的解释。2.1.

[[3]](" \l "_ftnref3" \o ")在训练开始时施加重量噪声会导致严重的不适。