计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300071

基于 BERT 的端到端语音合成方法

安 鑫 代子彪 李 阳 孙 晓 任福继
合肥工业大学计算机与信息学院 合肥 230601
合肥工业大学情感计算与先进智能机器安徽省重点实验室 合肥 230601
(xin.an@hfut.edu.cn)

摘 要 针对基于 RNN 的神经网络语音合成模型训练和预测效率低下以及长距离信息丢失的问题,提出了一种基于 BERT 的端到端语音合成方法,在语音合成的 Seq2Seq 架构中使用自注意力机制(Self-Attention Mechanism)取代 RNN 作为编码器。该方法使用预训练好的 BERT 作为模型的编码器(Encoder)从输入的文本内容中提取上下文信息,解码器(Decoder)采用与语音合成模型 Tacotron2 相同的架构输出梅尔频谱,最后使用训练好的 WaveGlow 网络将梅尔频谱转化为最终的音频结果。该方法在预训练 BERT 的基础上通过微调适配下游任务来大幅度减少训练参数和训练时间。同时,借助其自注意力(Self-Attention)机制还可以并行计算编码器中的隐藏状态,从而充分利用 GPU 的并行计算能力以提高训练效率,并能有效缓解远程依赖问题。与 Tacotron2 模型的对比实验表明,文中提出的模型能够在得到与 Tacotron2 模型相近效果的基础上,把训练速度提升 1 倍左右。

关键词:语音合成;循环神经网络;Seq2Seq;WaveGlow;注意力机制
中图法分类号 TP391

End-to-End Speech Synthesis Based on BERT

AN Xin,DAI Zi-biao,LI Yang,SUN Xiao and REN Fu-ji
School of Computer and Information,Hefei University of Technology,Hefei 230601,China
Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine,Hefei University of Technology,Hefei 230601,China

Abstract To address the problems of low training and prediction efficiency of RNN-based neural network speech synthesis models and long-distance information loss,an end-to-end BERT-based speech synthesis method is proposed to use the Self-Attention Mechanism instead of RNN as an encoder in the Seq2Seq architecture of speech synthesis. The method uses a pre-trained BERT as the model's Encoder to extract contextual information from the input text content,the Decoder outputs the Mel spectrum by using the same architecture as the speech synthesis model Tacotron2,and finally the trained WaveGlow network is used to transform the Mel spectrum into the final audio result. This method significantly reduces the training parameters and training time by fine-tuning the downstream task based on pre-trained BERT. At the same time,it can also compute the hidden states in the encoder in parallel with its Self-Attention mechanism,thus making full use of the parallel computing power of the GPU to improve the training efficiency and effectively alleviate the remote dependency problem. Through comparison experiments with the Tacotron2 model,the results show that the model proposed in this paper is able to double the training speed while obtaining similar results to the Tacotron2 model.

Keywords Speech synthesis,Recurrent neural network(RNN),Seq2Seq,WaveGlow,Attention mechanism

1 引言

语音合成(Speech Synthesis),又称文语转换(Text-to-Speech[1],TTS)技术,指计算机通过分析将任意文本转化为流畅语音。语音合成作为实现人机语音交互系统的核心技术之一[2],是语音处理技术中一个重要的方向,其应用价值越来越受到重视。作为人机语音交互的出口,语音合成的效果直接影响到人机交互的体验。一个高质量的、稳定的语音合成系统能够让机器更加地拟人化,使人机交互过程更加自然。

近年来,随着人工神经网络的迅速发展,语音合成研究的

到稿日期:2021-03-08 返修日期:2021-06-04 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金联合资助项目(U1613217);安徽省重点研究与开发计划项目(202004d07020004);中央高校基本科研业务专项资金(JZ2020YYPY0092)
This work was supported by the Joint Funds of the National Natural Science Foundation of China(U1613217),Key Research and Development Projects of Anhui Province of China(202004d07020004) and Fundamental Research Funds for the Central Universities of Ministry of Education of China(JZ2020YYPY0092).
通信作者:代子彪(1224269321@qq.com)

重心从基于参数拼接的方法转向了基于神经网络模型的方法[3]。诸多研究表明端到端的语音合成模型在语音合成领域取得了很好的效果。Wang 等提出的端到端的文语转化模型 Tacotron[4]能直接从文本预测产生梅尔频谱图，然后通过 Griffin-Lim[5]算法合成音频结果。该模型是一种序列到序列(Seq2seq)[6]的体系结构，用于从字符序列中产生幅度谱图，它是一个端到端的语音合成模型，输入为文本向量，输出为梅尔频谱图。Shen 等引入了 Tacotron2[7]，这是一个将文本转化为对应语种音频的神经网络架构。它主要包含两个部分：一部分是由 RNN 构成的序列到序列特征预测网络，该网络将输入文本数据映射到 Mel 标度图谱中；另一部分是一个声码器，该模型的声码器是一个修改后的 WaveNet[8]模型，它可以利用 Mel 频谱来合成语音音频。该模型在 MOS 评分中取得了 4.53 分，是当时的最好成绩。此模型结合了 Tacotron 和 WaveNet 各部分的优势，能合成高质量的音频。Arik 等提出了一个基于神经网络的端到端语音合成模型 Deep Voice1[9]，它可以将文本转化为目标语种的音频。随后，Arik 等在 Deep Voice1 的基础上二次迭代，提出了一种使用低维数据嵌入来增强文本到语音的训练的模型 Deep Voice2[10]，该模型可以产生不同语调的声音。Ping 等提出了一个基于 Seq2seq 架构的端到端语音合成模型 Deep Voice3[11]，该模型采用全卷积的方式实现了文本到 Mel 频谱的转化。

上述基于注意力机制[12-13]的 Seq2Seq 的编码器和解码器通常使用循环神经网络(Recurrent Neural Network，RNN)，比如 LSTM 以及 GRU 等。然而，RNN 作为一种自回归模型，其第 i 步的输入包含了第 $i-1$ 步输出的隐藏状态，这种时序结构限制了训练和预测过程中的并行计算能力。此外，这种结构还会导致当输入序列过长时，来自许多步骤之前的信息在传递过程中逐渐消失，进而使生成的上下文信息存在偏差。

为了解决上述问题，Aaron 等提出了一种并行训练模型 Parallel WaveNet[14]，他们引入了一种概率密度蒸馏的方法，从一个训练过的 WaveNet 中训练一个并行前馈网络。该方法整合了逆自回归流和波形网的特性，这些特性保证了 WaveNet 训练的有效性以及自回归流的有效采样。Prenger 等提出了一个能够从梅尔频谱图生成语音的基于流的网络 WaveGlow[15]。它结合了基于流的生成模型 Glow[16]的高效推理和预测以及热门语音合成模型 WaveNet 训练快、效果好的优点，可提供高质量音频的合成，并且无需进行自回归。

同样地，为了解决前文所述的两个问题，本文提出了一种使用 BERT[17]来替换 TTS 模型中编码器的新模型。该模型结合了 Tacotron2 和 BERT 的优势，使用预训练模型 BERT 作为编码器从输入的文本中提取上下文信息，并通过微调的方式来拟合下游任务。BERT 模型在 11 项 NLP 任务上取得的最佳成绩已经表明其具有很强的泛化能力和特征表示能力，其具有的自注意力机制可以同时计算输入序列的注意力相关性，在提高并行能力的同时缓解了 RNN 长距离信息丢失的问题。本文模型使用文本作为输入，输出为梅尔频谱图，并使用 WaveGlow 作为声码器来合成音频。WaveGlow 结合了 Glow 和 WaveNet 的优点，可以提供快速、高效和高质量的音频合成而不需要自回归。WaveGlow 只使用单个网络实现，只使用单个代价函数进行训练，这使得训练过程简单而稳定。

2 背景技术

本节首先介绍 Seq2Seq 模型，随后简要介绍 Tacotron2 和 BERT 模型。

2.1 Seq2Seq 模型

Seq2Seq 模型(Sequence to Sequence)也称为 Encoder-Decoder 模型，其架构如图 1 所示。该模型包含两个部分：编码器 Encoder 用于编码输入序列的信息，将任意长度的序列信息编码成固定长度的上下文向量；Decoder 是解码器，它从编码器输出的上下文向量中提取信息，并解码为输出序列。编码器和解码器常采用循环神经网络(RNN)的组合。该技术突破了传统的固定大小输入的框架，使得经典深度神经网络模型开始运用于翻译、文本自动摘要和机器人自动问答以及一些回归预测任务上，并被证实在 NLP 相关领域的应用中有着不俗的表现。

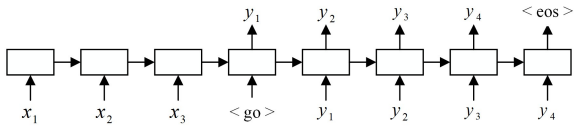


图 1 Seq2seq 模型架构

Fig. 1 System architecture of Seq2seq Model

Seq2Seq 模型将输入序列 $(x_1, x_2, \cdots, x_T)$ 转换为输出序列 $(y_1, y_2, \cdots, y_{T'})$，这一过程由编码输入与解码输出两个环节组成。前者负责把输入序列编码为一个固定长度的向量，将这个向量作为输入传给后者，输出可变长度的向量。并且每个时刻的输出 y_t 由先前时刻的输出 $(y_1, y_2, \cdots, y_{t-1})$ 共同决定。大多数情况下，输入序列长度与输出序列长度不相同，即 $T \neq T'$。在神经机器翻译(NMT)中，此转换基于条件概率 $p(y_1, y_2, \cdots, y_{T'} \mid x_1, x_2, \cdots, x_T)$ 将一种语言的输入语句转换为另一种语言的输出语句，即：

$$h_t = encoder(h_{t-1}, x_t)$$

$$s_t = decoder(s_{t-1}, y_{t-1}, c_t)$$

其中，h_t 和 s_t 分别表示编码器和解码器的隐藏状态，$\boldsymbol{c}_t$ 是输入序列通过注意力机制计算出来的上下文向量。

$$\boldsymbol{c}_t = attention(s_{t-1}, h)$$

$p(y_1, y_2, \cdots, y_{T'} \mid x_1, x_2, \cdots, x_T)$ 可以通过下方计算：

$$p(y_1, \cdots, y_{T'} \mid x_1, \cdots, x_T) = \prod_{t=1} p(y_t \mid y_1, \cdots, y_{t-1}, X)$$

$$p(y_t \mid y_1, \cdots, y_{t-1}, X) = softmax(f(s_t))$$

其中，$f(\cdot)$ 表示一个全连接层变换。对于翻译任务，$softmax$ 函数需要计算每个维度对应位置上词汇表中每个单词的概率。但是，在 TTS 任务中不需要 $softmax$ 函数，直接将解码器输出的隐藏状态输入到频谱图转换模型中。

2.2 Tacotron2

Tacotron2 是由 Google Brain 成员在 2017 年提出的一个基于神经网络的 End-to-End 语音合成框架，其架构如图 2 所示。它包含两个部分：声谱预测网络和声码器(Vocoder)。声谱预测网络是一个 Encoder-Attention-Decoder 网络，用于将输入的字符序列预测为梅尔频谱的帧序列。声码器是

WaveNet 的修订版，用于将预测的梅尔频谱帧序列产生时域波形（生成音频）。

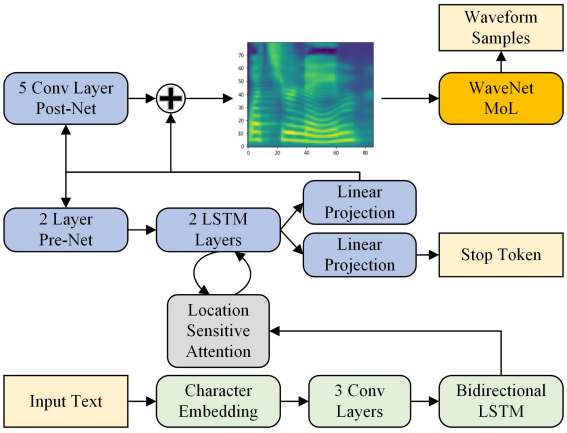


图 2　Tacotron2 架构

Fig. 2　System architecture of Tacotron2

输入文本序列在经过 Embedding 层后，先通过一个 3 层的 CNN 层提取长距离的上下文信息，然后将结果传递给由一个双向 LSTM 构成的 Encoder 层。上一时刻 $t-1$ 的梅尔频谱帧 y_{t-1}（推断时是预测帧，训练时是真实帧）首先使用 2 层全连接网络（2 Layer Pre-Net）进行处理，其输出与 $t-1$ 时刻输出的上下文向量 $\boldsymbol{c}_{t-1}$ 相加输入到 2 层的 LSTM 中。该输出将用于计算此时新的上下文向量 $\boldsymbol{c}_t$，并将其与 2 层 LSTM 的输出相连接，用来分别预测具有 2 个不同线性投影的梅尔频谱图帧和停止标志。最后将预测的梅尔频谱帧馈送到具有残差连接的 5 层 CNN 来进行精调。

2.3　预训练语言模型-BERT

BERT（Bidirectional Encoder Representations from Transformers）是基于 Self-Attention 的预训练语言模型，它完全免除了复杂性和重复性，其架构如图 3 所示。它是 Transformer[18] 模型的 Encoder 部分，由两个子网，即 Self-Attention 层和 Feed Forward 层构成，每个子网后接一个残差连接和归一化层。由于 BERT 内部存在自注意力机制，它可以并行地训练输入文本的双向深度表示；其次，可以在预训练好的 BERT 模型后接一个输出层，通过微调该层来适配下游任务，而不需要大量的特定于任务的额外训练，可以大大减少训练参数和训练时间。

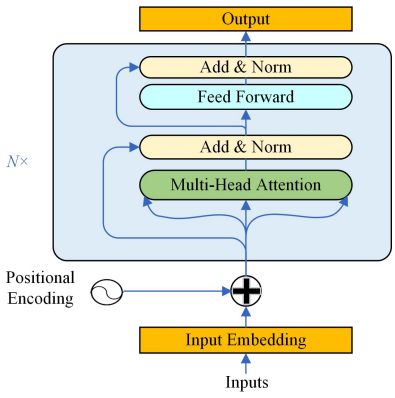


图 3　BERT 架构

Fig. 3　System architectureof BERT

相比原来的 RNN 以及 LSTM，BERT 可以做到并发执行，同时提取词在句子中的关系特征，并且能在多个不同层次提取关系特征，进而更全面地反映句子语义。BERT 在概念上很简单，在实际实验过程中得到的结果也很好。结果表明，在 11 种 NLP 任务上 BERT 都刷新了最好效果，并且在机器阅读理解顶级水平测试 SQuAD1.1 中的两个衡量指标上，全面超越了人类的表现。

在当下的 NLP 研究领域，随着计算机算力的不断增强，越来越多的通用语言表征的预训练模型（Pre-trained Models，PTMs）逐渐涌现出来，这对下游的 NLP 任务非常有帮助，可以避免从零开始训练新的模型。预训练模型有 3 个好处：1）预训练可以从大规模语料中学习到通用的语言表示，并用于下游任务；2）预训练提供了更优的模型初始化方法，有助于提高模型的泛化能力和加速模型收敛；3）预训练可以作为在小数据集上的一种避免过拟合的正则化方法[19]。

BERT 问世后，在 NLP 各个任务上都取得了非凡的结果。微调预训练好的 BERT 模型的方法可以在不同任务下泛化能力很好的原因，主要有如下 3 点：1）预训练可以使模型有一个好的初始化，相比随机初始化，其可以有相对较广的最优解范围（Wider Optima），并更快达到最优解（Easier Optimization）；2）由于 flat 和 wide 的 optima，训练损失函数以及泛化误差函数连续性好，不容易出现过拟合的情况，微调 BERT 可以产生更好的结果；3）BERT 的 lower layers 在微调过程中更难改变（Invariant），说明靠近输入的层可以学到更多的语言表征（Transferable Representations of Language）[20]。与基于 RNN 的模型相比，使用 BERT 作为解码器有 4 个优点：1）基于在大数据集上经过训练的模型，BERT 通过微调额外的输出层来适配下游任务，使得待训练参数量更少；2）Self-Attention 可以并行处理编码器的输入，不需要进行自回归计算，从而大大减少了计算时间，提升了训练效率；3）Self-Attention 机制可以同时从上下文中提取信息来建立长期依赖关系，从而避免传统 RNN 长距离信息丢失的问题；4）BERT 使用位置信息编码，使输入信号在前向和后向传播时，任意位置之间的间隔缩短到 1，这使得模型学习句子中词与词之间的关系、词与句子之间的关系时不再受限于位置，极大地提高了模型训练效率。

2.4　WaveGlow

尽管 Griffin-Lim 算法可以在不破坏左右相邻的幅度谱和自身幅度谱的情况下，求一个近似相位，但其会产生特有的人工痕迹并且合成的语音保真度较低，因此需要换成神经网络合成器来恢复信号的相位信息。

目前，大多数文本到语音的框架使用了 WaveNet 语音编码器来合成高保真的音频波形，但由于自回归采样太慢，其在实际应用中存在局限性。近期，研究者提出的 Parallel WaveNet[21] 通过整合逆向自回归流到并行采样中实现了实时的音频合成。然而，Parallel WaveNet 不仅需要两个阶段的训练，还需要一个训练良好的教师网络，并且如果只使用 probability distillation 来训练容易导致模式崩溃。因此，本文采用 WaveGlow 作为语音编码器来合成音频。

WaveGlow 是一个基于流的能够从梅尔频谱图生成高质量

的语音的网络。它结合了基于流的生成模型 Glow 的高效推理和预测以及热门语音合成模型 WaveNet 训练快、效果好的优点，提供了快速、高效和高质量的音频合成，而无需进行自回归。WaveGlow 仅使用单个网络来实现，而使用单个成本进行训练可以最大化训练数据的似然性，使训练过程简单、稳定。

3 基于 BERT 的语音合成模型

本节详细介绍了 BERT TTS 模型的架构，并分析了每个部分的功能。其总体架构如图 4 所示。

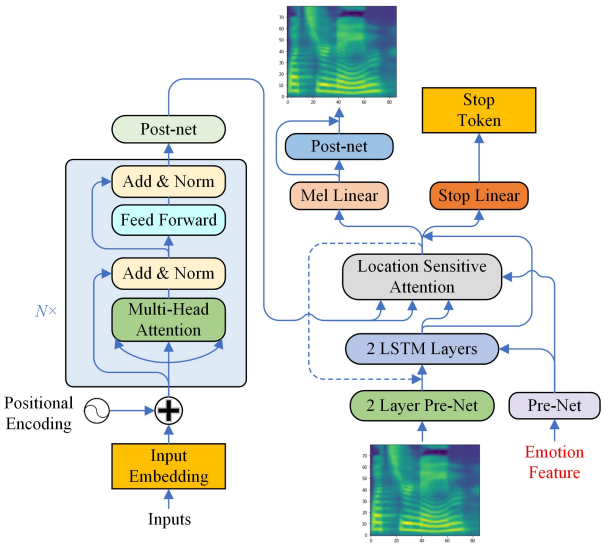


图 4 本文提出的模型架构

Fig. 4 System architecture of our model

3.1 Encoder

在 Tacotron2 中，编码器是双向 RNN，本文将使用 2.3 节的 BERT 对其进行替换。与原始的双向 RNN 相比，Multi-Head Attention 将注意力分散到多个子空间中，从而可以从多个不同方面建模输入序列之间的关系，并且直接在任意两个 token 之间建立长期依赖关系，因此每个 token 的输出都考虑了整个序列的全局上下文，这对合成音频韵律至关重要，尤其在句子较长的情况下，由此生成的样本听起来更加平滑、自然。BERT 支持的输入句子长度为 512 个字节，另外，采用 Multi-Head Attention 代替 RNN 可以通过并行计算来提高训练和预测速度。在 BERT 后接一个维度为 512 的全连接层后，处理网络(Post-Net)作为编码器的输出层，经过训练去适配下游任务。本文通过冻结 BERT 所有层的参数，将预训练模型 BERT 作为一个特征提取器，只重新训练其后接的 Post-Net 层的参数，得到新的权重。

文献[22]指出，原版 BERT 基于 WordPiece 的分词方式，会把一个完整的词切分成若干个子词，在生成训练样本时，这些被分开的子词会随机被 mask。比如 playing 在基于 WordPiece 分词时会被分割为 play 和 ing，如果此时在随机 mask 时，ing 被选中，则会导致训练效果不佳。随后 Google 推出了全词 Mask(Whole Word Masking，WWM)升级版本。在全词 Mask 中，如果一个完整的词的部分 WordPiece 子词被 mask，那么同属该词的其他部分也会被 mask，即当 ing 被选中 mask 时，会同时将 play 也包含进去。

本文选取的是 Google 开源的包含 WWM 的英文预训练模型 BERT-large，其层数 N 为 24，输出层隐藏状态维度为 1024。

3.2 Decoder

与 Tacotron2 类似，解码器是一个自回归网络，采用具有 location-sensitive 注意力机制的 2 层 RNN，可以一次从一帧的编码输入序列中预测梅尔频谱帧。上一个时刻解码器的输出先通过一个预处理网络(Pre-Net)，其中包含 2 个全连接层，每层有 256 个 ReLU 隐藏单元。实验表明，作为信息瓶颈的预处理网络对于注意力的学习至关重要；然后，将预处理网络的输出和注意力网络的输出连接起来并通过两个具有 1024 个单元的单向 LSTM 层；LSTM 层的输出和注意力上下文向量的连接通过线性变换进行投影，以预测目标梅尔频谱帧；最后将预测的梅尔频谱图通过 5 层卷积后处理网络，并对结果进行残差连接以提高整体效果，每层后处理网络由 512 个形状为 5×1 的卷积核组成，并进行 Batch-Normalization，然后在除最后一层之外的所有层上用 tanh 函数激活。

3.3 Mel Linear 和 Stop Linear

本文使用两个不同的线性层分别预测梅尔频谱图和结束标志位。值得一提的是，对于停止线性层而言，每个序列的末尾只有一个正样本，表示“停止”，而其他帧则有数百个，即负样本有数百个。这可能导致模型无法正确预测结束帧从而陷入无线循环中，Tacotron2 即存在这个问题。本文在计算二进制交叉熵损失时，在尾部的停止标记上施加正权重(5.0～8.0)，从而有效地解决了该问题。比如使用 Tacotron2 模型生成“Hello Hello”这句话的音频结果，会得到一个 11 s 的音频，而本文提出的模型生成的音频符合预期。

4 实验评估

为了测试模型音频建模的性能，本文设计了 Tacotron2 与本文模型的对比实验。本节首先介绍本文所使用的数据集和实验环境，然后介绍本文所使用的评估方法以及实验结果。为了评估模型性能，本文做了主观配对测试，通过平均意见得分(Mean Opinion Score，MOS)来评估合成的音频质量。在主观配对比较测试中，测试者听完每个合成结果，会对合成语音的自然度进行 5 分制打分(1 为很差，5 为很好)，打分间隔为 0.5，对不同模型不同训练阶段的样本进行测试。

4.1 数据集

本文所有的实验都基于 LJSpeech-1.1 语音数据集。该数据集包含 13100 个单声道演讲者的短音频片段，这些片段来自 7 本非小说类书籍。该数据包含在家庭环境中使用内置麦克风在 MacBook Pro 上记录的大约 24 h 的语音数据。

4.2 实验设置

特征提取和预处理：本文使用的语音信号采样率为 22050 Hz，采样位为 16 bit，使用 Hamming 窗处理，帧长为 50 ms，帧移为 12.5 ms，预加重系数为 0.97。所有数据集中的句子都经过了预处理，去除了音频前后的空白部分，以及对文本进行了标准化处理(例如将“9”转写成“nine”)。

本文实验均采用多 GPU(2 个 2080Ti，11 GB)并行训练，并且均使用了半精度浮点数(FP16)加速训练。本文的训练

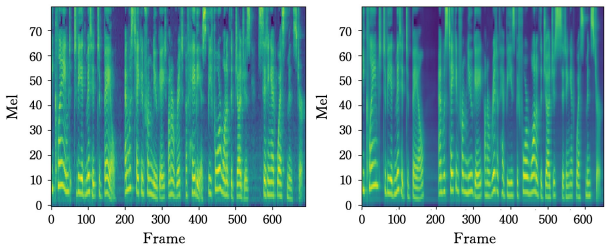
过程包括使用 BERT 作为编码器对特征预测网络进行单独训练,以及在相同训练集上单独训练 WaveGlow 来合成高质量音频。

为了训练特征预测网络,本文在单个 GPU 上执行 batch-size 为 64 的极大似然训练程序(在训练时,解码器传递给下一时刻的不是预测值而是真实值),并且冻结了在大数据集上预训练好的 BERT 模型参数,使其作为一个特征提取器从输入文本中提取上下文信息。本文在 BERT 后接一个 Post-Net 模块,它是包含一个隐藏层的全连接网络。隐藏层的隐藏单元数目与输入单元维度一样,为 512。该层采用 ReLu(Rectified Linear Units)激活函数,并使用 0.9 的 dropout 进行正则化处理。本文使用 Adam[23] 优化器,其中 $\beta_1=0.9$,$\beta_2=0.999$,$\varepsilon=10^{-8}$,并且将学习率固定为10^{-3}。本文还应用权重为10^{-6}的 L_2 正则化。

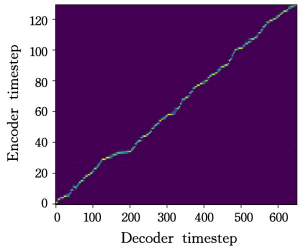
在训练 WaveGlow 时,本文使用原始音频的频谱图作为 WaveGlow 网络的输入。对于 WaveGlow,本文使用带有 librosa mel 过滤器默认设置的 80 个 bin 的梅尔频谱图,即每个 bin 通过过滤器长度进行归一化,并且刻度与 HTK 相同。梅尔频谱图的参数是:FFT 大小为 1 024,跳数为 256,窗口大小为 1 024。

4.3　特征预测情况

图 5(a)、图 5 (b)给出了句子"He claimed to be admitted to bail,and was taken from Newgate on a writ of habeas before one of the judges sitting at Westminster."的真实的梅尔声谱图和预测的梅尔声谱图。通过观察可以看到,合成的梅尔声谱图和原音频十分接近,此外,经过训练,模型也学习到了清晰平滑的特征网络参数(见图 5(c)),这表明该模型在处理频谱图细节方面表现较好。



(a)Real mel spectrogram　(b)Predicted mel spectrogram



(c)Attention alignment

图 5　特征预测

Fig. 5　Feature prediction

4.4　评估

本文从内部数据集中选取了 10 句(非训练集内)具有不同长度的样本作为固定的评估集,在保证文本内容一致并排除其他干扰因素的前提下,评估不同模型不同训练阶段生成的这 10 个句子(包括真人录音)的 MOS 得分。参与打分的志愿者共 34 人,主要是来自本校英语专业的研究生,其中包括 6 名母语为英语的志愿者。实验的详细结果如表 1 所列。

表 1　MOS 评分的实验结果

Table 1　Result of MOS score

Epochs	System	MOS	CMOS
500	Tacotron2	3.22	—
	Ours	3.39	—
1 000	Tacotron2	3.90	0.00
	Ours	3.90	0.02
—	Ground Truth	4.54	—

从表 1 可以看出,在迭代 500 次时本文模型合成的音频效果明显优于 Tacotron2,这主要是因为本文模型的收敛速度比 Tacotron2 更快。而在相同实验环境下迭代 1 000 个epochs 后,本文模型获得了与 Tacotron2 相同的 MOS 评分。因此,本文进一步采用相对评价意见打分(The Comparison Mean Option Score,CMOS)来对比 Tacotron2 模型和本文模型的性能。在 CMOS 的测试实验中,测试人员每次会听到两种音频(分别由本文模型和 Tacotron2 合成),并使用[-3,3]之间的整数打分,以此来评估前后两个音频的主观差距。可以看到,本文模型以 0.02 的优势获得了更高的评分。

同时,为了分析输入文本长度对模型性能的影响,本文对比了不同长度文本的音频合成结果的 MOS 得分情况。首先将上文选取的 10 个句子按长度递增编号为 1-10,每个句子的 MOS 得分如图 6 所示。由于 BERT 最长支持 512 维的输入(不足 512 维进行补 0 操作),不存在 RNN 长距离信息丢失的问题,在长音频合成中,本文模型的效果要明显优于 Tacotron2。当句子较长时,随着句子长度的增加,Tacotron2 模型的合成效果逐渐变差,而本文提出的模型 BERT TTS 则不存在这一问题。

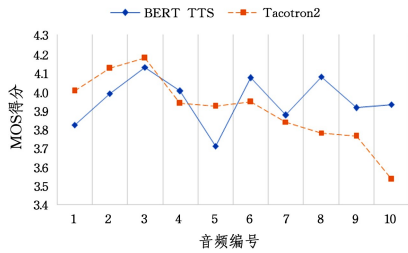


图 6　不同长度的语音合成样本 MOS 的比较

Fig. 6　Comparison of MOS for speech sample in different lengths

4.5　训练时间对比

在训练过程中,本文通过插桩的方式记录了每个训练 step 的耗时,并且每间隔 4 个训练步骤计算一次训练损失。在本文的实验环境下(2 个 GPU,FP16 启用,batch-size 为 64),把本文模型的单个训练 step 耗时均值约为 0.35s,比相同实验环境下的 Tacotron2(约为 0.65s)耗时缩短了近一半。

训练过程中,损失函数的变化曲线如图 7 所示。多次实验结果表明,两个模型的损失函数最终都会收敛到 0.3 附近。从图 7 可以看出,本文模型比 Tacotron2 的收敛速度更快,即本文模型使用更少的训练步骤就可以得到效果较好的音频结果,大大节约了实验耗时。

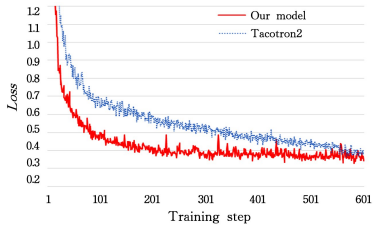


图 7 训练时损失变化

Fig. 7 Changes in training loss

本文采用 BERT 作为编码器，它内部的 Self-Attention 结构可以并行地计算编码器的输出，因此可以节省大量的训练时间。并且由于本文模型是基于预训练好的 BERT 微调来适配下游任务，因此收敛速度比 Tacotron2 更快。

结束语 本文提出了一个基于 Tacotron2 和 BERT 的神经网络模型，并进行了部分修改使得该模型能适应 TTS 任务。本文模型能充分利用 GPU 的并行计算能力，从而获得更快的训练速度和预测速度。并且，该模型能从输入序列中获取远距离信息，使其在长文本语音合成中的效果比 Tacotron2 等模型更好。实验结果表明，本文模型能够在得到与 Tacotron2 模型相近效果的基础上，把训练速度提升 1 倍左右。此外，在长文本语音合成的场景下，本文模型的效果要明显优于 Tacotron2。从实际使用过程来看，本文模型的稳定性更好，不存在类似 Tacotron2 在预测过程中偶尔无法正确预测结束帧的问题。

虽然本文使用 BERT 作为 TTS 模型的 Encoder，但是 Decoder 仍然是自回归的，采用文献[24]提出的 Transformer 作为编码器和解码器是后期的研究内容之一。此外，如何改进当前模型使其可以更好地契合中文语音合成任务也是作者后面工作的一个方向。

参 考 文 献

[1] TAYLOR P. Text-to-speech synthesis [M]. New York: Cambridge University Press, 2009.

[2] FUNG P, SCHULTZ T. Multilingual spoken language processing[J]. IEEE Signal Processing Magazine, 2008, 25(3): 89-97.

[3] PAN X Q, LU T L, DU Y H, et al. Overview of Speech Synthesis and Voice Converrsion Technology Based on Deep Learning [J]. Computer Science, 2021, 48(8): 200-208.

[4] ZHANG B, QUAN C Q, REN J F. Overview of Speech Synthesis in Development and Methods[J]. Journal of Chinese Computer System, 2016, 37(1): 186-192.

[5] WANG Y, SKERRY-RYAN R, STANTON D, et al. Tacotron: toward end-to-end speech synthesis [J]. arXiv: 1703. 10135, 2017.

[6] GRIFFIN D, LIM J S. Signal estimation from modified short-time Fourier transform[J]. 1984 IEEE Transactions on Acoustics Speech and Signal Processing, 1984, 32(2): 236-243.

[7] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems, 2014, 27: 3104-3112.

[8] SHEN J, PANG R, WEISS R J, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]// Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 4779-4783.

[9] OORD A V D, DIELEMAN, ZEN H, et al. WaveNet: a generative model for raw audio[J]. arXiv: 1609. 03499, 2016.

[10] ARIK S O, CHRZANOWSKI M, COATES A, et al. Deep voice: Real-time neural text-to-speech[J]. arXiv: 1702. 07825, 2017.

[11] GIBIANSKY A, ARIK S O, DIAMOS G F, et al. Deep Voice 2: Multi-Speaker Neural Text-to-Speech[C]// Proceedings of the Advances in 2017 Neural Information Processing Systems. United states: NIPS. 2017: 2963-2970.

[12] PING W, PENG K, GIBIANSKY A, et al. Deep voice 3: Scaling text-to-speech with convolutional sequence learning[J]. arXiv: 1710. 07654, 2017.

[13] CHRORWSKI J K, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition [J]. Advances in Neural Information Processing Systems, 2015, 28: 577-585.

[14] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409. 0473, 2014.

[15] OORD A, LI Y, BABUSCHKIN I, et al. Parallel wavenet: Fast high-fidelity speech synthesis[C]// Proceedings of the International Conference on Machine Learning. Cambridge MA: JMLR, 2018: 3918-3926.

[16] PRENGER R, VALLE R, CATANZARO B. Waveglow: A flow-based generative network for speech synthesis[C]// Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 3617-3621.

[17] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1×1 convolutions[C]// Proceedings of the Advances in 2018 NeuralInformation Processing Systems. United states: NIPS, 2018: 10215-10224.

[18] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810. 04805, 2018.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the Advances in 2017 Neural Information Processing Systems. United states: NIPS, 2017: 5998-6008.

[20] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: A survey[J]. arXiv: 2003. 08271, 2020.

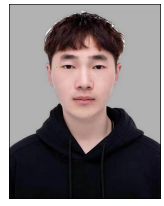
[21] HAO Y, DONG L, WEI F, et al. Visualizing and understanding the effectiveness of BERT[J]. arXiv: 1908. 05620, 2019.

[22] PING W, PENG K, CHEN J. Clarinet: Parallel wave generation in end-to-end text-to-speech[J]. arXiv: 1807. 07281, 2018.

[23] CUI Y, CHE W, LIU T, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[J]. arXiv: 2004. 13922, 2020.

[24] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv: 1412. 6980, 2014.

AN Xin, born in 1987, Ph. D, associate professor, is a member of China Computer Federation. His main research interests include embedded systems and machine learning.



DAI Zi-biao, born in 1994, postgraduate. His main research interests include machine learning and affective computing.

（责任编辑：柯颖）