文章编号：2095-2163(2021)02-0169-05 中图分类号：TP183 文献标志码：A

一种 RNN-T 与 BERT 相结合的端到端语音识别模型

郭家兴，韩纪庆
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 端到端语音识别模型由于结构简单且容易训练，已成为目前最流行的语音识别模型。然而端到端语音识别模型通常需要大量的语音-文本对进行训练，才能取得较好的识别性能。而在实际应用中收集大量配对数据既费力又昂贵，因此其无法在实际应用中被广泛使用。本文提出一种将 RNN-T（Recurrent Neural Network Transducer，RNN-T）模型与 BERT（Bidirectional Encoder Representations from Transformers，BERT）模型进行结合的方法来解决上述问题，其通过用 BERT 模型替换 RNN-T 中的预测网络部分，并对整个网络进行微调，从而使 RNN-T 模型能有效利用 BERT 模型中的语言学知识，进而提高模型的识别性能。在中文普通话数据集 AISHELL-1 上的实验结果表明，采用所提出的方法训练后的模型与基线模型相比能获得更好的识别结果。

关键词： 语音识别；端到端模型；BERT 模型

An end-to-end speech recognition model combining RNN-T and BERT

GUO Jiaxing, HAN Jiqing
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 The end-to-end speech recognition model has become one of the most popular speech recognition models due to its simple structure and easy training. However, it usually needs a large number of speech-text pairs for the training of an end-to-end speech recognition model to achieve a better performance. In practical applications, it is very laborious and expensive to collect a large number of the paired data, resulting in the model cannot be widely used. This paper proposes a method of combining the Recurrent Neural Network Transducer (RNN-T) model with the Bidirectional Encoder Representations from Transformers (BERT) model to solve the above problems. It replaces the prediction network part in the RNN-T with the BERT model and fine-tunes the entire network, thus the RNN-T model effectively uses linguistic information to improve model recognition performance. The experimental results on the Chinese mandarin data set AISHELL-1 show that, compared with the baseline system, the system using the proposed expansion method achieves better recognition results.

【Key words】 speech recognition; end-to-end model; BERT model

0 引 言

近年来，各种基于深度神经网络的端到端模型在语音识别（Automatic Speech Recognition，ASR）领域正逐渐成为研究热点。不同于传统的语音识别模型，端到端模型不再需要将输入语音帧和给定文本标签进行一一对齐，其仅包含一个单独的序列模型，可以直接将输入的语音特征序列映射为识别的文本序列，简化了识别的过程。同时模型不依赖语言模型和发音词典，降低了对专家知识的要求[1-3]。目前，端到端语音识别模型主要包括基于注意力机制的编解码模型[4-5]、连接时序分类（Connectionist Temporal Classification，CTC）模型[6-7]、基于循环神经网络转换器（Recurrent Neural Network Transducer，RNN-T）的模型[8-9]三种。其中，RNN-T模型是由 Graves 等人针对 CTC 的不足所提出的改进方法。相比于 CTC，RNN-T 可以同时对输入和输出序列的条件相关性进行建模，而且对输入和输出序列的长度没有限制。这使得 RNN-T 模型更加适合语音任务，因此本文拟围绕 RNN-T 模型来展开研究工作。

时下的大量研究表明[10-14]，端到端语音识别模型仍然存在着语料资源有限所导致训练不充分等一系列问题。而收集大量语音-文本对非常困难，这导致端到端语音识别模型在实际应用中的表现欠佳。最近的工作表明，可以使用纯文本数据来改善其性能。文献[5]用词级语言模型组成 RNN 输出网格，文献[8]用外部语言模型对搜索算法进行重新打分。文献[15-16]在波束搜索期间合并了字符级语言模型，而文献[17]采用知识迁移的方法，先

基金项目： 国家重点研发项目(2017YFB1002102)。

作者简介： 郭家兴(1995-)，男，硕士研究生，主要研究方向：语音识别；韩纪庆(1964-)，男，博士，教授，博士生导师，主要研究方向：语音信号处理、音频信息处理。

收稿日期： 2020-09-28

对大规模外部文本训练语言模型,再将该语言模型中的知识迁移到端到端语音识别系统中。这些方法在解码阶段将端到端模型与其它语言模型结合在一起,可以有效改善语音识别模型的性能,但是都需要额外的步骤来集成和微调单独的语言模块,因此都不是真正意义上的端到端模型。

为了解决上述问题,同时考虑到 BERT(Bidirectional Encoder Representations from Transformers)模型[18]是目前对语言学信息建模最好的模型,本文提出一种将 RNN-T 模型与 BERT 模型进行联合优化的方法,就可以高效利用 BERT 模型所提供的语言学信息,也是一种真正的端到端模型。

1 提出方法

1.1 RNN-T 模型及其局限性分析

1.1.1 基于 RNN-T 的端到端语音识别模型

基于 RNN-T 的端到端语音识别模型能够很好地将声学信息和语言学信息进行联合优化,在端到端语音识别任务中取得了目前最好的性能,通常由 3 部分构成:编码器(Encoder)、预测网络(Predict Network)和联合网络(Joint Network)。其中,编码器的功能就类似于传统语音识别系统的声学模型,通过将输入的声学特征序列转化为发音基元序列,预测网络给出对应的语言学信息,联合网络的作用是结合语言学信息和发音基元序列产生对应的转录文本,整个模型结构如图 1 所示。

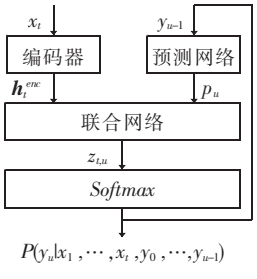

图 1 RNN-T 模型结构[9]

Fig. 1 RNN-T model structure[9]

假设输入的声学特征序列为 $X=[x_1,x_2,\cdots,x_T]$,标注文本序列为 $Y=[y_0,y_1,\cdots,y_{U-1}]$。研究中将文本标签进行扩展 $Y'=[eos,y_0,y_1,\cdots,y_{U-1}]$,这里的 *eos* 表示开始的字符。针对每一时刻的输入输出对 (x_t,y_u),RNN-T 模型的计算过程如下:

$$\boldsymbol{h}_t^{enc}=Enc(\boldsymbol{x}_t)\ , \tag{1}$$

$$\boldsymbol{p}_u=Pre(\boldsymbol{y}_{u-1})\ , \tag{2}$$

$$\boldsymbol{z}_{t,u}=Joint(\boldsymbol{h}_t^{enc},\boldsymbol{p}_u)\ , \tag{3}$$

$$\boldsymbol{y}_u=Softmax(\boldsymbol{z}_{t,u})\ . \tag{4}$$

其中,$Enc(\cdot)$ 表示编码器网络;$\boldsymbol{h}_t^{enc}\in R^m$ 表示当前时刻编码网络的声学特征;$Pre(\cdot)$ 表示预测网络;$\boldsymbol{p}_u$ 表示预测网络在当前时刻的上下文向量;$Joint(\cdot)$ 表示联合网络,主要负责将输入的文本特征和声学特征进行融合;$\boldsymbol{z}_{t,u}$ 表示网络当前时刻的预测向量,结合了声学特征以及文本特征;$\boldsymbol{y}_u$ 表示网络的预测输出结果。

编码器一般使用长短期记忆(Long Short-Term Memory, LSTM)网络[19],将输入语音帧 $\boldsymbol{x}_t$ 转换为高级声学表示 $\boldsymbol{h}_t^{enc}$,类似于基于 CTC 的标准语音识别器中的声学模型。因此,与在 CTC 中一样,编码器网络的输出 $\boldsymbol{h}_t^{enc}$ 取决于先前语音帧 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t$ 的序列。预测网络中的 RNN 保证模型能同时考虑输入和输出序列的时序关系,从而使 RNN-T 消除了 CTC 中的条件独立性假设。具体地说,预测网络接收最后一个非空白标签 $\boldsymbol{y}_{u-1}$ 作为输入,以产生输出 $\boldsymbol{p}_u$。研究表明,联合网络是一个前馈网络,能将预测网络和编码器的输出组合起来以产生 $logit(\boldsymbol{z}_{t,u})$,此后将经过一个 *Softmax* 层以产生分布输出目标。

RNN-T 模型不仅解决了 CTC 中输出之间的条件独立性假设,以及缺少语言建模能力的不足,还使用了共同建模的思路来对语言模型和声学模型进行联合优化;同时,模型具有在线解码等诸多优点,是一种比较有前景的模型。因此,本文首先搭建基于 RNN-T 结构的端到端语音识别基线模型。

1.1.2 RNN-T 模型的局限性分析

RNN-T 模型也存在不足。一方面,由于在 RNN-T 模型中,声学建模与语言学建模已被整合在一个网络中,其仅用一个目标函数进行优化,这就要求训练数据必须同时包含输入和输出序列。然而在实际应用中配对数据的获取十分困难。另一方面,RNN-T 模型并不能像 CTC 一样与传统的 WFST 结合,在第一遍解码中,未能利用大型语言模型的好处,而 RNN-T 的预测网络所提供的上下文信息,只能在一定程度上缓解这种劣势。

实际上传统的语音识别模型也会出现上述问题。传统语音识别模型结构如图 2 所示。由图 2 可知,在传统语音识别模型中,通常采用独立的声学模型和语言模型分别建模声学信息和语言学信息。首先,使用声学模型去识别每一个发音基元,将输入的声学特征序列转化为发音基元序列;然后,在发音词典和语言模型的帮助下,通过搜索算法在发音基元序列中得到一条最佳路径,这条最佳路径就对应了

识别的转录文本序列。对于容易出错的词，语言模型没有见过或者很少见过这种搭配，导致搜索算法计算出的概率得分很低。所以要提高语音识别模型的识别准确率，就必须重新扩充语言模型部分，旨在使模型对容易出错的词也能计算出一个比较高的概率得分。因此传统的语音识别模型可以利用比训练集的转录文本多几个数量级的纯文本数据，来单独训练语言模型部分，以更新语言学的知识，从而保持声学模型部分不动。然而，通过扩充语言模型的方式并不适用于 RNN-T 模型，因为在 RNN-T 模型中训练数据和扩充数据都必须是平行的文本和语音对。

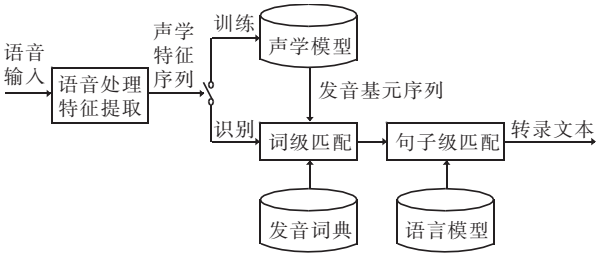

图 2　传统语音识别模型结构图

Fig. 2　Traditional speech recognition model structure diagram

1.2　用 BERT 模型替换预测网络

根据 1.1 节中的分析，RNN-T 模型在实际应用中表现不好是因为缺乏训练数据，进而导致模型的语言学信息建模不充分。而 RNN-T 的预测网络所提供的上下文信息，只能在一定程度上缓解这种劣势。鉴于传统语音识别方法可以直接用大量文本数据单独训练语言模型部分，从而扩充模型的语言学信息，在 RNN-T 模型中，编码器部分相当于声学模型，预测网络相当于语言模型。参考传统语音识别方法的经验，直观有效的方法就是对预测网络进行扩充。因此，本文提出使用更强大的语言模型来替换 RNN-T 模型的预测网络部分，以在推理时提供更具表示性的语言学信息。

BERT 模型是目前对语言学信息建模最好的语言模型[20]，与其它语言模型不同，BERT 采用双向语言模型的方式，能够更好地融合上下文的信息。同时，预训练的 BERT 模型在实际使用时，只需要根据具体任务额外加入一个输出层进行微调即可，而不用为特定任务来修改模型结构。本文使用 BERT 模型来替换 RNN-T 模型的预测网络部分，使联合网络在进行解码的过程中，通过 BERT 模型引入外部的语言学信息来进行辅助解码。网络结构如图 3 所示。替换后的模型在进行解码时，由预测网络提供当前时刻的上下文向量变为由 BERT 模型提供对应信息。

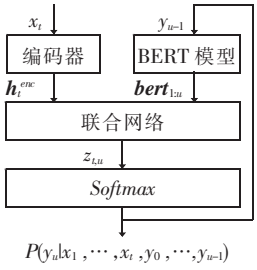

图 3　结合 BERT 的 RNN-T 模型

Fig. 3　RNN-T model combined with BERT

假设当前时刻的输入输出对为（$\boldsymbol{x}_t$，$\boldsymbol{y}_u$）。在上一时刻，联合网络将预测得到的非空白标签 $\boldsymbol{y}_{u-1}$ 送给 BERT 模型，BERT 模型将其与前期预测得到的所有非空白标签 $y_0, y_1, \cdots, y_{u-2}$ 拼接为矩阵［$y_0, y_1, \cdots, y_{u-1}$］作为输入，并输出一个固定长度的向量 $\boldsymbol{bert}_{1:u}$，代表了 BERT 模型根据已解码出的句子提供的上下文向量。相比于预测网络只根据最后一个非空白标签 y_{u-1} 预测得到的上下文向量 $\boldsymbol{p}_u$，$\boldsymbol{bert}_{1:u}$ 中引入了长时的上下文信息，同时因为 BERT 是用大量文本数据预训练过的，其所提供的 $\boldsymbol{bert}_{1:u}$ 也就具有更强的表示性。此后，联合网络接收 $\boldsymbol{bert}_{1:u}$ 以预测当前时刻的 $\boldsymbol{z}_{t,u}$，式(3)变为式(5)，也就是：

$$\boldsymbol{z}_{t,u} = Joint(\boldsymbol{h}_t^{enc}, \boldsymbol{bert}_{1:u}). \tag{5}$$

1.3　微调 RNN-T 模型

1.2 节中介绍的将 BERT 模型与 RNN-T 模型进行结合的方法，通过使用 BERT 模型替换 RNN-T 模型的预测网络部分，实现了在推理时利用 BERT 模型提供的语言学信息。

然而实验结果表明，直接替换的方法会导致模型的识别性能下降，这是因为 BERT 没有参与训练，只是在 RNN-T 模型进行解码时提供相应信息，从而导致了 BERT 模型和 RNN-T 的编码器部分不匹配。例如，$t-1$ 时刻联合网络预测的字符为“新”，而 BERT 模型预测下一个字符是“冠”，但语料库中并没有这个词，这就导致联合网络没有见过 BERT 模型提供的信息，从而出现错误。

解决方法是微调 RNN-T 模型。具体来说，就是在用 BERT 模型替换掉 RNN-T 的预测网络部分后，再用训练语料库重新训练一遍整个模型。在这个过程中 BERT 模型参与了训练，使联合网络逐渐适应 BERT 模型提供的信息，进而使编码器和 BERT 模型相互匹配。

2 实验与结果分析

2.1 实验数据

实验基于 2 种普通话语料库:AISHELL-1[21]和 AISHELL-2[22]。其中,AISHELL-1 包含 180 h 语音数据,AISHELL-2 包含 1 000 h 语音数据。使用 Kaldi 提取 40 维的 FBank 特征,每个特征都被重新调整为在训练集上具有零均值和单位方差。

在实验中,本文使用 AISHELL-1 训练 RNNT 模型,将 AISHELL-2 的转录文本作为文本数据集,训练 BERT 模型。

2.2 模型结构和实验设置

在基线 RNN-T 模型中,编码器由 5 层双向长短时记忆(Bidirectional Long Short-Term Memory, BLSTM)网络组成,每层有 700 个单元,正向和反向各有 350 个单元。预测网络由 700 个门控循环单元(Gated Recurrent Unit, GRU)的单层组成,联合网络结合了声学和语言学信息,由 700 个单元的单向前馈网络组成,使用 tanh 作为激活函数。

在实验设置方面,模型采用声学特征作为输入,标注文本作为输出序列,实现端到端的语音识别模型;模型直接进行解码,以提取输出字符序列,而无需使用单独的发音模型或外部语言模型;采用字错误率(Character Error Rate, CER)作为语音识别效果的评价指标。

2.3 实验结果与分析

本文的实验结果见表 1。RNN Transducer 是使用 AISHELL-1 数据集训练的基线模型。RNN Transducer * 模型是用 BERT 模型替换 RNN-T 模型中的预测网络部分,并在推理时提供语言学信息的结果,可以发现字错误率大幅度上升。这是因为 BERT 模型并没有参与训练,只是在 RNN-T 模型解码时提供相应信息,导致 BERT 模型和 RNN-T 的编码器部分不匹配。RNN Transducer + Bert 是用 AISHELL-1 数据集对整个模型进行重训练的结果,相当于对联合网络进行微调,使编码器部分与 BERT 模型之间相互匹配。与基线模型比较后可知,本文提出的方法相对降低了 5.2% 的字错误率,提高了模型的识别性能。

表 1 在 AISHELL-1 数据集上的实验结果

Tab. 1 Experimental results on the AISHELL-1 data set

模型	数据集	CER/%
RNN Transducer	AISHELL-1	11.5
RNN Transducer *	AISHELL-1	16.3
RNN Transducer + Bert	AISHELL-1	10.9

3 结束语

本文针对基于 RNN-T 的端到端语音识别模型,提出了一种与 BERT 模型进行结合的方法。该方法通过用 BERT 模型替换 RNN-T 中的预测网络部分,对整个网络进行微调,从而使 RNN-T 模型在训练和解码过程中能够有效利用 BERT 提供的语言学信息,进而提高模型的识别性能。最后,在 AISHELL 中文普通话数据集上对所提出的方法进行了评估,实验结果表明,该方法能够获得更好的 ASR 性能。

参考文献

[1] 韩纪庆, 张磊, 郑铁然. 语音信号处理[M]. 2 版. 北京:清华大学出版社, 2013.

[2] ALTER. 语音识别进化简史:从造技术到建系统[J]. 大数据时代, 2019(9):50-59.

[3] PRABHAVALKAR R, RAO K, SAINATH T N, et al. A comparison of sequence-to-sequence models for speech recognition[C]//Interspeech. Stockholm, Sweden: dblp, 2017: 939-943.

[4] GRAVES A, GOMEZ F. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning. New York, USA: ACM, 2006: 369-376.

[5] MIAO Y, GOWAYYED M, METZE F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Dammam: IEEE, 2015: 167-174.

[6] GRAVES A. Sequence transduction with recurrent neural networks[J]. arXiv preprint arXiv:1211.3711, 2012.

[7] RAO K, SAK H, PRABHAVALKAR R. Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer[C]//2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Okinawa, Japan: dblp, 2017: 193-199.

[8] CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 4960-4964.

[9] BAHDANAU D, CHOROWSKI J, SERDYUK D, et al. End-to-end attention-based large vocabulary speech recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 4945-

4949.
[10] KARITA S, WATANABE S, IWATA T, et al. Semi-supervised end-to-end speech recognition [C]//Interspeech. Hyderabad, India: dblp, 2018: 2-6.
[11] BASKAR M K, WATANABE S, ASTUDILLO R F, et al. Self-supervised Sequence-to-sequence ASR using unpaired speech and text[C] //Interspeech. Graz, Austria: dblp, 2019: 3790-3794.
[12] RENDUCHINTALA A, DING S, WIESNER M, et al. Multi-modal data augmentation for end-to-end ASR[C]//Interspeech. Hyderabad, India: dblp, 2018: 2394-2398.
[13] HORI T, ASTUDILLO R, HAYASHI T, et al. Cycle-consistency training for end-to-end speech recognition [C]// ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 6271-6275.
[14] HAYASHI T, WATANABE S, ZHANG Yu, et al. Back-translation-style data augmentation for end-to-end ASR[C]// 2018 IEEE Spoken Language Technology Workshop (SLT). Athens: IEEE, 2018: 426-433.
[15] MAAS A, XIE Z, JURAFSKY D, et al. Lexicon-Free conversational speech recognition with Neural Networks [C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Colorado, USA: ACL, 2015: 345-354.
[16] HORI T, WATANABE S, ZHANG Yu, et al. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM [C]//Interspeech. Stockholm, Sweden: dblp, 2017: 949-953.
[17] BAI Ye, YI Jiangyan, TAO Jianhua, et al. Learn spelling from teachers: Transferring knowledge from language models to sequence-to-sequence speech recognition [C]// Interspeech. Graz, Austria: dblp, 2019: 3795-3799.
[18] DEVLIN J, CHANG Mingwei, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.
[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780.
[20] JIANG D, LEI X, LI W, et al. Improving transformer-based speech recognition using unsupervised pre-training [J]. arXiv preprint arXiv: 1910.09932, 2019.
[21] BU Hui, DU Jiayu, NA Xingyu, et al. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline [C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, South Korea: IEEE, 2017: 1-5.
[22] DU Jiayu, NA Xingyu, LIU Xuechen, et al. AISHELL-2: Transforming mandarin ASR research into industrial scale [J]. arXiv preprint arXiv: 1808.10583, 2018.

（上接第 168 页）

5　结束语

本文提出将树莓派和 NodeMCU 开发板联合开发一款性价比极高的智能语音交互家居系统，设计使用树莓派 4B+在 Linux 系统下运行百度云平台提供的 SDK 加上麦克风阵列和 CSI 摄像头实现语音识别、语音合成、人脸检测等主要功能，实现人与硬件设备语音交互。NodeMCU 通过内置 ESP8266 芯片使用 Arduino IDE 编译将传感器采集数据通过 TCP 协议与物联网云平台连接，实现数据的远程存储、家庭环境信息的远程观测、远程控制。设计采用 2 款主流硬件，树莓派实现完整的操作系统安装，调用物联网云平台 API 连接，实现家居环境监测、家用电器控制、安防管理及语音交互功能。产品将传感器技术、WiFi 技术、物联网平台、LabVIEW 技术结合起来实现对环境实时监控并及时控制，采用无线技术，不需任何布线，实施方便。系统稳定性，可靠性好，设计系统软硬件拓展方便，安全性高。整个系统操作简单，方便，能在短时间内进行熟练操作，产品功耗低，可以用移动电源供电，安装快捷方便，能快速组建一个实时远控系统。

参考文献

[1] 周宏伟. 智能家居的系统结构及相关无线通信技术研究[J]. 数字通信世界，2019(3)：115.
[2] 满莎，杨恢先，彭友，等. 基于 ARM9 的嵌入式无线智能家居网关设计[J]. 计算机应用，2010，30(9)：2541-2544.
[3] 刘硕，赵彦博，杜佳林，等. 基于蓝牙的物联网智能家居系统设计[J]. 通信与信息技术，2020(02)：72-73，61.
[4] 孙全宝. 基于语音识别的智能家居系统设计[J]. 物联网技术，2020，10(7)：105-106，110.
[5] 薛辉. 基于语音合成的智慧导游系统的研究与设计[J]. 信息技术，2020，44(2)：112-115，120.