文章编号：2095-2163(2021)03-0076-05 中图分类号：N33 文献标志码：A

基于 NAO 机器人的 BLSTM-CTC 的声学模型研究

胡希颖，王大东，陈佳欣
（吉林师范大学 计算机学院，吉林 四平 136000）

摘 要：针对于 NAO 机器人自身语音识别准确率低的问题，提出一种基于 NAO 机器人的 BLSTM-CTC 的声学模型研究方法。基于 BLSTM-CTC 的声学模型进行建模，以 BLSTM 为声学模型和 CTC 为目标函数，以音素作为基本建模单元，建立中文语音识别端到端系统。实验结果证明，本文算法相较于 NAO 机器人自身，取得了良好识别效果。

关键词：语音识别；BLSTM-CTC；NAO

Acoustic model of BLSTM-CTC based on NAO robot

HU Xiying, WANG Dadong, CHEN Jiaxin
(College of Computer, Jilin Normal University, Siping Jilin 136000, China)

【Abstract】 Aiming at the problem of low accuracy of NAO robot's own speech recognition, an acoustic model research method based on NAO robot BLSTM-CTC is proposed. Based on the acoustic model of BLSTM-CTC, an end-to-end system for Chinese speech recognition is established by taking BLSTM as the acoustic model and CTC as the objective function, and taking phonemes as the basic modeling unit. Experimental results show that compared with NAO robot itself, the proposed algorithm achieves good recognition performance.

【Key words】 speech recognition; BLSTM-CTC; NAO

0 引 言

语音识别是语音信号处理领域的一项重要研究内容，其中的基于深度学习的识别方法则在近年来引起了学界的广泛关注[1]。基于深度学习的识别方法是利用神经网络来构建模型、训练数据，并已取得了较好的识别效果，现正广泛应用于智能家居以及相关的学术研究等领域。作为备受学界瞩目的智能机器人，NAO 本身自带语音识别模块，但却因受到自身处理速度和存储能力的限制，识别效果一般。考虑到 NAO 机器人自身的软硬件资源较为有限，只依靠 NAO 自身来提高语音识别准确率的难度较大。基于此，本文即提出以了 BLSTM[2] 为声学模型和 CTC 为目标函数，利用 WFST 进行解码，对模型结构进行训练和学习，并将其移植到 NAO 机器人上，从而获得更好的识别结果，提升机器人的学习能力。

1 模型结构

LSTM（Long Short - Term Memory）最早由 Hochreiter & Schmidhuber 在 1977 年提出[3]，后经 Alex Graves 完善并获得广泛应用[4]。LSTM 主要由 2 部分组成。一个是传统的外部 RNN 循环；一个是内部精致的“门”结构，包括 sigmoid 神经网络层和按位乘法操作。LSTM 的“门”分别是输入门、输出门、遗忘门，3 个门控单元控制和保护 cell 的信息到细胞状态[5]。LSTM 基本结构如图 1 所示。

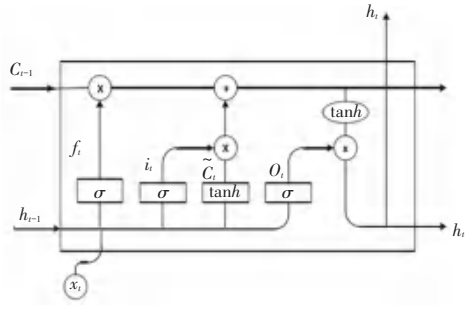


图 1 LSTM 的基本结构图

Fig. 1 Basic structure diagram of LSTM

图 1 中，遗忘门 f 决定从细胞状态 cell 中遗弃哪些数据信息。其对应数学公式可写为：

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f, \quad (1)$$

其中，σ 为 sigmoid 函数，读取 h_{t-1} 和 x_t，并输出到每个细胞状态 C_{t-1} 中一个 0 ~ 1 之间的数值。

细胞状态 cell 确定可存放信息数据，输入门 i_t，确定信息的更新与否，并在 tan h 层创建新的候选向量 $\tilde{C}_t$，如此则用新的主语来更新代替旧的细胞状态。

LSTM 只注重关联上一历史时间段的信息建模，单向地考虑历史信息对后续信息的影响。因此，

作者简介：胡希颖（1995-），女，硕士研究生，主要研究方向：人工智能。

收稿日期：2021-02-15

当需要获取全文信息时，就要引入双向循环神经网络（Bidirectional Recurrent Neural Network，Bi-RNN），在处理连续数据的基础上，不仅可以学习正向规律 $\overrightarrow{h}$，也可学习反向规律 $\overleftarrow{h}$，将正反向网络传播相结合来获取相对完整信息，提高识别拟合度。Bi-RNN 的基本结构如图 2 所示。

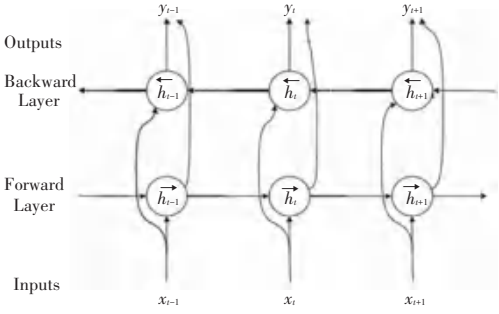


图 2　Bi-RNN 的基本结构图

Fig. 2　Basic structure diagram of Bi-RNN

在隐层中增加了一层反向传播序列后，Bi-RNN 的隐层计算中由前向序列 $\overrightarrow{h}$ 和后向序列 $\overleftarrow{h}$ 组成，在 t 时刻的 LSTM 更新方式如下所示：

$$\overrightarrow{h}_t = W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}, \tag{2}$$

$$\overleftarrow{h}_t = W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}, \tag{3}$$

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_o. \tag{4}$$

2　基于连接时序分类的语音识别系统

采用传统神经网络训练声学模型方法时，先是根据声学模型的基本单元进行建模，在训练时还需使用 GMM 与标签进行对齐，并将目标函数作为训练标准。本文用 BLSTM-CTC 系统在训练声学模型时采用端到端的训练方式，不同于传统的混合方法基于 eesen 框架的 RNN 使用基于交叉熵（CE）准则训练帧级标签，而是采用 CTC 函数学习帧与序列的对齐，并使用 WFST 进行解码[6]，BLSTM-CTC 系统结构如图 3 所示。

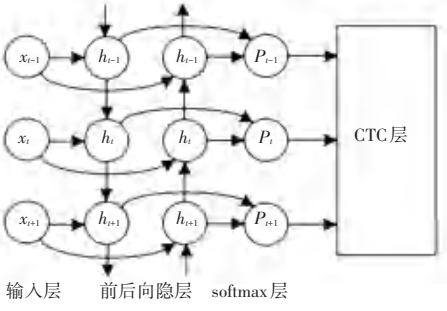


图 3　BLSTM-CTC 语音识别系统结构图

Fig. 3　Structure diagram of BLSTM-CTC speech recognition system

2.1　连接时序分类 CTC 技术

CTC（Connectionist Temporal Classification）技术作为目标函数无需强制预先对齐输入与输出帧级别信息，可直接对标签和语音特征之间的映射进行建模。RNN 中 softmax 层的输出序列、即 CTC 层的输入，softmax 层中的 k 个节点与 CTC 中训练数据的标签序列一一对应；对未输出的标签也需建模，在此基础上，增加一个单元（blank）。假定长度是 T 的输入序列 x，输出向量 y_t，在 t 时刻 softmax 分类层输出音素或空白的概率 k 表示为：

$$P_r(k,t \mid x) = \frac{\exp(y_k^t)}{\sum_{k'}\exp(y_t^{k'})}, \tag{5}$$

CTC 经过学习后得到由音素和 blank 组成的标注序列 a 的输出路径概率为：

$$P_r(a \mid x) = \prod_{t=1}^{T} P_r(a,t \mid x), \tag{6}$$

由于标注的重复性和 blank 插入的影响，音频序列与转录后去掉空白标签的路径具有多重对应关系，因此，输入序列 x 对应的输出标签概率为：

$$P_r(y \mid x) = \sum_{a\in\beta^{-1}} P_r(a \mid x), \tag{7}$$

其中，$a \to y$ 的映射获取 β，β 的逆过程表示为 β^{-1}，映射过程把空白类去除的同时将重复序列合并得到 y 目标函数，即：

$$CTC(x) = -\log P(y \mid x), \tag{8}$$

通过已知的输入序列找到最大概率的输出路径、即 CTC 网络解码的最佳路径为：

$$a^* = \mathrm{argmax}_a p(a \mid x). \tag{9}$$

CTC 路径求和随着输入序列的增加，计算复杂度越来越增大，为解决这一实际问题，在输出序列 z 的首尾及每对输出标签序列之间插入索引是“0”的 blank 标签，从而将得到的增广式扩充标签序列 $l = (l_1,\cdots,l_{2U+1})$ 用于语音识别中前后向算法（Forward-backward Algorithm）计算路径似然估计[7]。

标签序列 z 的似然估计概率计算如下：

$$P(z \mid X) = \sum_{u=1}^{2U+1} \alpha_t^u \beta_t^u, \tag{10}$$

其中，t 为 1 到 T 时刻中的任意一帧。CTC 目标函数 $\ln\Pr(z \mid X)$ 对 RNN 网络输出 y^t 求微分，则 $\ln\Pr(z \mid X)$ 相对于 y_t^k 的一阶导为：

$$\frac{\partial \ln\Pr(z \mid X)}{\partial y_t^k} = \frac{1}{\Pr(z \mid X)} \cdot \frac{1}{y_t^k} \sum_{u\in\gamma(l,k)} \alpha_t^u \beta_t^u. \tag{11}$$

由式（11）可见，目标函数可进行微分，所以 b^t、b^i、b^o、b^f 在求导过程中误差影响可以忽略，RNN 在接收 softmax 层反向传播过程中即可更新参数。

2.2 WFST 解码

一般情况下,应用于 CTC 训练输出模型的解码方法均有些不足。一是不能把单词级语言模型进行有效的整合[8];二是只能在特定约束条件下进行集合[9],因此需要高效解码。本文基于发声特点将语言模型、词典和 CTC 输出用 WFST 进行编译,建立一个基于 WFST 的搜索图实现高效完整性的解码操作。WFST 实质上是一个 FSA(Finite - state Acceptor),相应的每个转换都包含输入符号、输出符号和权重[10]。

WFST 解码由 3 个部分组成,分别是:标记(Token)、语法(Grammar)和词典(Lexicon)。对此拟做阐释分述如下。

(1)语法 G:基于语言模型 n-gram 编码了符合语法的单词序列。初始节点用节点 O 表示,每个边的权重即当前对应字或词的概率。

(2)标记 T:编码了语音 CTC 标签序列 L 到词典单元 L' 的一对多的映射关系 $\phi(l)$。在词典单元中,帧级别标签序列进行 WFST 存在空白标签 Φ 和重复序列,例如处理五帧后的 RNN 可能存在的标记序列“AAAAA”、“ΦΦAAΦ”、“ΦAAΦΦ”,token 的 WFST 可把上述三种序列均映射为一个“A”的词典单元。

(3)词典 L:WFST 将标签序列 L 的词序列映射到字序列进行编码。空的输入和输出用<eps>表示。

3 个独立的 WFST 在编译后,把语法 G 和词典 L 进行组合获得 LG 网络,再通过确定化和最小化算法针对 LG 网络进行处理,同时减少搜索图的占用和优化 WFST 网络,最终结合 CTC 标签生成完整的搜索图,也就是:

$$S = T \circ \min(\det(L \circ G)). \tag{12}$$

在搜索图 S 中。T、min、det 分别表示组合、最小化和确定化操作[11]。S 通过编码将获取的 CTC 标签映射到字序列,此方法较 HMM 模型 CTC 解码速度和性能均大幅度提高。

3 实验结果与分析

3.1 实验数据集

本节的基于 NAO 机器人的 BLSTM-CTC 声学模型研究是基于清华大学开源的 THCHS-30 中文数据集。该数据集是由 50 人录制的、共计时长为 35 h 的声音数据,数据中的采样率和量化位数分别为 16 kHz 和 16 bit。其中,训练集占 74.7%,共 10 000句;开发集占 6.7%,共 893 句;测试集则占 18.6%,共2 495句,并且每个集合之间均不存在相同录制人。语言模型为 3-gram 模型。

3.2 实验设置

本次实验中的硬件配置是 Ubuntu Linux 操作系统和 NAO 机器人的麦克风;实验软件配置是搭建 TensorFlow1.5 框架结构和 Python2.7 编程语言。实验中搭建的基于 BLSTM-CTC 端到端语音识别系统,输入特征参数 MFCC 帧长为 256,Mel 频率倒谱系数为 26,每个时间段有 494 个 MFCC 特征数,语音输入的窗函数选用汉明窗。

3.3 实验结果分析

端到端系统建模能力强于基线系统,但不同的网络隐藏层数对系统性能的影响也存在差异性。表 1给出了不同的隐藏层数,即 2 层、3 层和 4 层之间系统的 *WER* 值对比。

表 1 不同网络层数下的 BLSTM-CTC 系统性能对比

Tab. 1 Comparison of BLSTM-CTC system performance under different network layers

LSTM 层	遗忘门偏置	*WER*/ %
2	1	26.07
3	1	25.06
4	1	28.35

由表 1 可知,LSTM 网络层数为 3 层时,相较于 2 层和 4 层,系统的 *WER* 值分别降低了 1.01% 和 2.28%。当网络层由 2 层丰富到 3 层时,结构得到完善,性能获得提升;当网络层由 3 层增加到 4 层时,由于训练语料库的短缺,导致网络欠拟合,反而抑制系统准确率的提升。因此,3 层的网络系统结构最优。

本实验的网络模型结构是由 3 层全连接层网络作为输入,每层包含 1 024 个节点,设置最佳学习率为 0.001,共进行 120 次迭代,每次迭代共循环 267 次,每次取 8。训练中选取句子字数相同、但循环次数不同的 3 组数据进行对比,分别是 69、139 和 209,每次迭代训练后均对训练损失、错误率和训练时间进行输出。以音素为基本单元进行建模,输出层激活函数是 softmax 函数,其输出标签数为 47,其中包含一个静音标签和 blank 标签以及 45 个音素。

文中选取前 22 次的迭代数据,分析 3 种不同循环次数进行对比,如图 4 所示。随着迭代次数的增加,在端到端语音识别系统中循环 69 次的正确率峰值最大;循环 139 次相较其他两者识别变化更加稳定;循环 209 次初始错误率最低。可见循环次数越多,错误率越小。

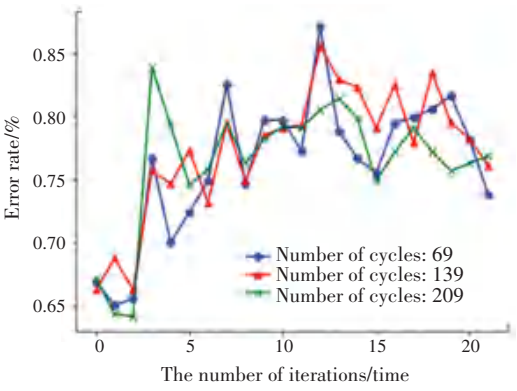


图 4　比较不同循环次数 BLSTM-CTC 语音识别 *WER* 对比

Fig. 4　*WER* comparison of BLSTM-CTC speech recognition with different cycle times

不同循环次数 BLSTM-CTC 语音识别损失对比如图 5 所示。由图 5 可知，端到端语音识别系统循环次数 69 次时，初始损失为 304.81，较其他两者损失相比过大；当循环次数为 209 次时，初始损失则为 292.24，当迭代数目增加时，损失均呈现逐渐下降趋势，不同次数间的损失数值变化区别不明显，可见循环次数越小损失变化越明显。综上可知，循环次数为 209 时，损失变动小，鲁棒性更强。

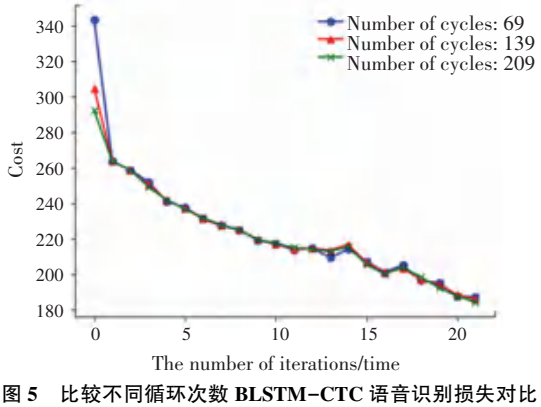
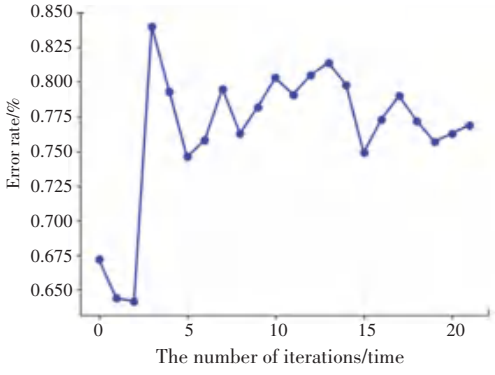


图 5　比较不同循环次数 BLSTM-CTC 语音识别损失对比

Fig. 5　Comparison of BLSTM-CTC speech recognition losses with different cycle times

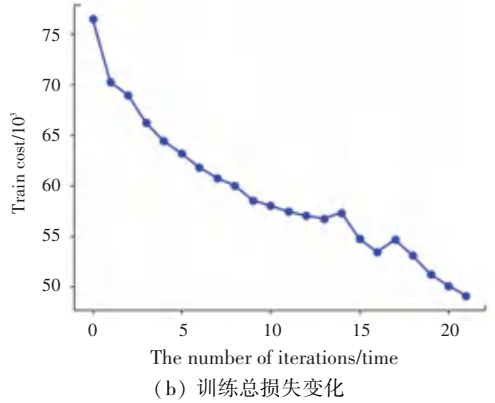
BLSTM-CTC 语音识别 *WER* 和损失变化则如图 6 所示。由图 6 可知，随着迭代次数的变化，训练损失大幅度降低，错误率变化不稳定，但趋势处于降低状态，最终的识别准确率为 74.4%。实现 NAO 机器人语音识别鲁棒性的有效提高。

NAO 机器人、端到端系统对比见表 2。表 2 中，针对 NAO 机器人自身和使用端到端系统二者进行对比，依据词错误率（Word Error Rate，*WER*）作为评判标准。与最初的 NAO 识别准确率相比，BLSTM-CTC 系统将 *WER* 值降低 6.57%。研究中发现 *WER* 值成功降低，但仍存在一些不足，BLSTM-CTC 系统训练后不受外界附加条件影响和制约，但训练时间长。由此可见，两者鲁棒性均获得大幅度提高，但也都存在一定的弊端，因此，两者可相互弥补在不同的硬件配置条件下的不足，通过多种方案均可有效改善 NAO 机器人 *WER* 值。



（a）BLSTM-CTC 语音识别 *WER* 的变化

（a）Changes in BLSTM-CTC speech recognition *WER*



（b）训练总损失变化

（b）Change in total training loss

图 6　BLSTM-CTC 语音识别 *WER* 和损失变化图

Fig. 6　Variation of *WER* and loss in BLSTM - CTC speech recognition

表 2　NAO 机器人、端到端系统对比

Tab. 2　NAO robot and end-to-end system comparison

识别方法	训练时间/h	*WER*/ %
NAO 机器人	0	31.63
BLSTM-CTC	113.26	25.06

4　结束语

本文使用基于 BLSTM-CTC 的声学模型进行建模，建立了中文语音识别端到端系统，应用于 NAO 机器人。实验结果证明，使用端到端系统比 NAO 机器人自身的 *WER* 有了进一步的改善，为 NAO 机器人的语音处理领域提供了更多的思路。

参考文献

[1] 戴礼荣，张仕良，黄智颖. 基于深度学习的语音识别技术现状与展望[J]. 数据采集与处理，2017，32(2)：221-231.

（下转第 84 页）