文章编号:1671 - 6833(2018)05 - 0063 - 05

基于改进的 LSTM 深度神经网络语音识别研究

赵淑芳, 董小雨

(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要: 当前基于 LSTM 结构的神经网络语言模型中,在隐藏层引入了 LSTM 结构单元,这种结构单元包含一个信息储存较久的存储单元,对历史信息有良好的记忆功能. 但 LSTM 中当前输入信息的状态不能影响到输出门最后的输出信息,对历史信息的获取较少. 针对以上问题,笔者提出了基于改进的 LSTM (long short-term memory)网络模型建模方法,该模型增加从当前输入门到输出门的连接,同时将遗忘门和输入门合成一个单一的更新门. 信息通过输入门和遗忘门将过去与现在的记忆进行合并,可以选择遗忘之前累积的信息,使得改进的 LSTM 模型可以学到长时期的历史信息,解决了标准 LSTM 方法的缺点,具有更强的鲁棒性. 采用基于改进的 LSTM 结构的神经网络语言模型,在 TIMIT 数据集上进行模型测试,结果表明,改进的 LSTM 识别错误率较标准的 LSTM 识别错误率降低了 5%.

关键词: 长短时记忆(LSTM); 深度神经网络; 语音识别

中图分类号: TP39 **文献标志码:** A **doi:**10.13705/j.issn.1671 - 6833.2018.02.004

0 引言

N-Gram 语言模型是一种简单有效的统计语言模型,因其简单、易用、高效,在实际应用中一直发挥重要作用. N-Gram 语言模型将词看作离散的符号,容易出现数据稀疏问题. 随着 N-Gram 语言模型阶数的增长,需要估计的参数量会急剧增加并存在数据稀疏问题,无法有效利用长距离的上下文信息,且无法捕捉词与词之间的相似性[1].

基于神经网络语言模型的连续空间建模方法首先基于分布式假设条件,通过投影矩阵将离散的词映射到连续空间,形成相应的词矢量(word embedding)特征,以此为基础,将分布式的上下文矢量信息输入到神经网络,并在输出层预测下一个词出现的概率[2-3].

递归神经网络语言模型输入层不仅包含了当前输入,而且加入了当前词的全历史词信息. 递归神经网络语言模型理论上可以考虑无限个历史词信息,但是随着不断引入新词,存在记忆衰退严重问题,针对这一问题,有学者提出了基于 LSTM 结构的神经网络语言模型. LSTM-DNN 语言模型在隐藏层引入了 LSTM 结构单元[4-5],这种结构单元包含将信息储存较久的存储单元,这个记忆单元被一些特殊的门限保护[6-7],可以选择性地记忆网络误差回传参数,对历史信息有良好的记忆功能.

由于 LSTM 中当前 CEC 的状态不能影响到输出门的最后输出信息,所以笔者增加从当前 CEC 到输出门的连接更好地控制每一内存单元输出多少信息,将遗忘门和输入门合成一个单一的更新门. 信息通过输入门和遗忘门将过去与现在的记忆进行合并,长短时记忆神经网络可以选择遗忘之前累积的信息,这样改进的 LSTM 模型可以学到长时期的历史信息,具有更强的鲁棒性. 笔者采用基于改进的 LSTM 方法,解决了 LSTM 方法的缺点,在 TIMIT 数据集上,错误率降低了 5%.

1 神经网络

人工神经网络(artificial neural network, ANN)是一种应用类似大脑神经突触链接的结构进行信息处理的数学模型,由大量的节点(神经元)和相互之间的加权连接构成. 每个节点代表一种特定的输出函数,称为激励函数(activa-

收稿日期:2017 - 05 - 26;**修订日期:**2017 - 08 - 28

基金项目:国家自然科学基金资助项目(61202163);"十二五"山西省科技重大专项资助项目(20121101001);山西省教研项目(J2017078)

作者简介:赵淑芳(1978—),河北唐山人,太原科技大学副教授,主要从事图像处理、人工智能和软件工程等研究,E-mail:zsf166@163.com.

tion function)[8-9]. 每两个节点间的连接都代表一个通过该连接信号的加权值,称为权重(weight),这相当于神经网络的记忆. 网络的输出则根据网络的连接方式、权重值和激励函数的不同而不同.

人工神经网络特别适合因果关系复杂的非确定性推理、判断、识别和分类等问题. 可以通过预先提供的一批相互对应的输入输出数据,分析掌握两者潜在的规律,最终根据这些规律,用新的输入数据来推算输出结果[10]. 人工神经元是一个多输入/单输出的非线性元件,其模型如图 1 所示.

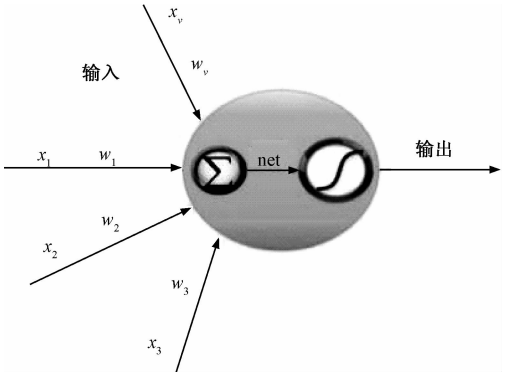

图 1 人工神经元模型

Fig. 1 Artificial neuron model

2 LSTM 深度学习基本思想

2.1 标准 LSTM 基本思想及模型构建

相比前馈型神经网络 DNN,循环神经网络(recurrent neural network,RNN)在隐层上增加了一个反馈连接,可以通过循环反馈连接看到前面所有时刻的信息,RNN 通过反馈连接将之前信息的记忆保留在中间的隐藏节点中,影响网络的输出[11]. 在传统的循环神经网络中,参数训练使用随时间进行反向传播(backpropagation through time,BPTT)算法,假设循环神经网络在每个时刻 t 都有一个监督信息,损失为 J_t,则整个序列的损失为:

$$\sum_{t=1}^{T} J_t. \tag{1}$$

损失 J 关于 U 的梯度可以用链式法则得到,

$$\frac{\partial J}{\partial U} = \sum_{t=1}^{T}\sum_{k=1}^{t}\frac{\partial h_k}{\partial U} f(h_i)\frac{\partial y_t}{\partial h_t}\frac{\partial J_t}{\partial y_t},$$

$$f(h_i) = \prod_{i=k+1}^{t} U^{\mathrm{T}}\,\mathrm{diag}(f'(h_{i-1})). \tag{2}$$

定义

$$\gamma = \| U^{\mathrm{T}}\,\mathrm{diag}(f'(h_{i-1}))\|. \tag{3}$$

公式(3)即为 γ^{t-k},若 $\gamma < 1$,当 $(t-k)\to\infty$ 时,$\gamma^{t-k}\to 0$ 会出现梯度消失. 当传播的时间比较长时,网络权重更新缓慢,参数训练时梯度需要随着时间进行反向传播,存在梯度爆炸和消失问题,无法体现 RNN 的长期记忆效果,因此需要一个存储单元来存储记忆,LSTM 模型由此被提出.

长短时记忆网络(LSTM network)又称长短时记忆循环神经网络(LSTM RNN),在深度学习中使用比较广泛.

(1)在长短时记忆神经网络(long short-term memory neural network,LSTM)中,引入一组称为记忆单元(memory units)的循环连接子网络来替换传统网络中的隐层节点. 用一个记忆细胞来进行线性的反馈传递.

(2)引入门机制(gating mechanism)控制信息的累积速度,提供对记忆细胞的写、读和重置操作,遗忘门 f_t 控制每一个内存单元需要遗忘多少信息,用来选择忘记过去某些信息[12]. 输入门 i_t 控制每一个内存单元加入多少新的信息,用来记忆现在的某些信息. 输出门 O_t 控制每一个内存单元输出了多少信息[13].

信息通过输入门和记忆门将过去与现在的记忆进行合并,输出门最后输出信息. LSTM 通过"控制门"的结构来去除或者增加信息到细胞状态[14],让信息选择式通过. 长短时记忆神经网络可以选择遗忘之前累积的信息,这样 LSTM 模型可以学到长时期的历史信息.

在 t 时刻,记忆单元 C_t 记录了到当前时刻为止的所有历史信息,并受 3 个"门"控制:输入门 i_t,遗忘门 f_t 和输出门 O_t. 3 个门的元素值在[0,1]之间. 在 t 时刻 LSTM 的更新方式如下:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1}), \tag{4}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1}), \tag{5}$$

$$O_t = \sigma(W_o x_t + U_f h_{t-1} + V_f c_{t-1}), \tag{6}$$

$$\overline{C_t} = \tan\mathrm{h}(W_c x_t + U_c h_{t-1}), \tag{7}$$

$$C_t = f_t \Theta C_{t-1} + i_t \Theta \overline{C}_t, \tag{8}$$

$$h_t = O_t \Theta \tan\mathrm{h}(C_t). \tag{9}$$

标准长短时记忆网络就是一种包含许多扩展记忆块的循环神经网络模型. 长短时记忆网络模型如图 2 所示.

2.2 改进的 LSTM 方法的基本思想

笔者在改进长短时记忆结构的基础上,将改进的结构应用于语音识别系统中,与基于 LSTM 模型的语音识别系统进行对比[15].

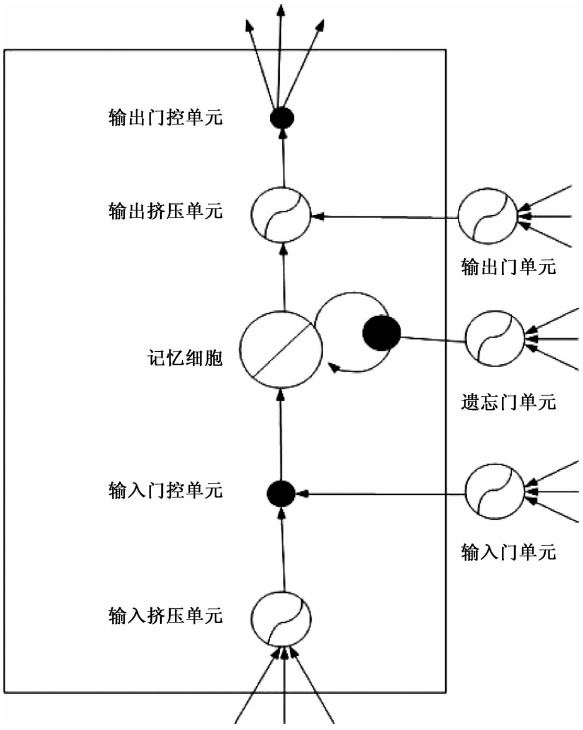

图 2　长短时记忆神经网络模型

Fig. 2　Neural network model for long and short duration memory

LSTM 中,当前输入信息的状态不能影响输出门的输出信息,增加当前输入门到输出门的连接,更好地控制每一内存单元输出多少信息.

将遗忘门和输入门合成了一个单一的更新门,耦合遗忘和输入门限由之前的 LSTM 分开确定什么忘记、添加什么新的信息变为一同做出决定. 在输入的时候才进行遗忘,在遗忘某些信息时才将新值添加到状态中[16]. 长短时记忆神经网络可以选择遗忘之前累积的信息,这样改进的 LSTM 模型可以学到长时期的历史信息,具有更强的鲁棒性. 改进的长短时记忆神经网络模型 LSTM 如图 3 所示.

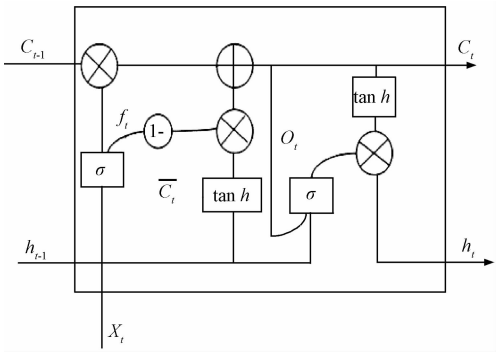

图 3　改进的长短时记忆神经网络模型 LSTM

Fig. 3　Improved short duration memory neural network model LSTM

在 t 时刻 LSTM 的更新方式如下:

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f), \tag{10}$$

$$O_t = \sigma(W_o[C_t, h_{t-1}, X_t] + b_0), \tag{11}$$

$$\overline{C_t} = \tanh[W_c \cdot [h_{t-1}, x_t] + b_c], \tag{12}$$

$$C_t = f_t \cdot C_{t-1} + (1\text{-}f_t) \cdot \overline{C_t}, \tag{13}$$

$$h_t = O_t \Theta \tanh(C_t). \tag{14}$$

2.3　改进的 LSTM 训练过程

改进的 LSTM 递归神经网络参数训练使用 BPTT 算法.

设在 t 时刻,网络的输入向量为 $x(t)$,隐含向量是 $h(t)$,网络的输出向量是 $o(t)$. 用 V 表示输入层与隐含层之间的连接权值,U 表示隐含层与隐含层之间的连接权值,W 表示隐含层与输出层之间的连接权值,b 和 a 分别表示隐含层和输出层的偏置.

$h^l(t)$表示第 l 个样本在 t 时刻的隐含变量;$o^l(t)$表示第 l 个样本在 t 时刻的输出变量;$\delta^l(v^l(t))$表示第 l 个样本在 t 时刻输出层的误差反向信号变量;$\delta^l(u^l(t))$表示第 l 个样本在 t 时刻隐含层的误差反向信号向量,$\frac{\partial L_N}{\partial W}$、$\frac{\partial L_N}{\partial V}$ 和 $\frac{\partial L_N}{\partial U}$ 分别表示对权值 W、V 和 U 的偏导;$\frac{\partial L_N}{\partial a}$、$\frac{\partial L_N}{\partial b}$ 分别表示对偏置 a 和 b 的偏导.

首先随机初始化所有的权值和偏置,初始化 $\frac{\partial L_N}{\partial W}, \frac{\partial L_N}{\partial V}, \frac{\partial L_N}{\partial U}, \frac{\partial L_N}{\partial a}, \frac{\partial L_N}{\partial b} = 0$.

随着时间 t 从 1 到 T 正向传播,对第 l 个样本在 t 时刻的隐含变量和输出变量进行定义. 在 $t=0$ 时刻,定义隐含变量都为 0,随着时间的展开,更新样本在 t 时刻的隐含变化量和输出变量.

$$h^l(t) = f(u^l(t)) = f(Vx^l/(t) ++ Uh^l(t-1) + b), \quad (1 \leqslant t \leqslant T);$$

$$o^l(t) = g(u^l(t)) = f(Wh^l(t) + a). \tag{16}$$

随着时间 t 从 T 到 1 反向传播,计算第 l 个样本在 t 时刻输出层的误差反向信号变量和隐含层的误差反向信号变量. 对权值 w、v、u 和偏置 a、b 的偏导进行更新.

$$\partial^l(v^l(t)) = (o^t) - y^l(t) \cdot g'(v^l(t)), \tag{17}$$

$$\partial^l(u^l(t)) = [(W)^{\mathrm{T}}\partial^l(v^l(t))] \cdot f'(u^l(t)), \tag{18}$$

$$\frac{\partial L_N}{\partial W} = \frac{\partial L_N}{\partial W} + \sum_{t=1}^{N} \delta^l(v^l(t))\,(h^l(t))^{\mathrm{T}}, \tag{19}$$

$$\frac{\partial L_N}{\partial V} = \frac{\partial L_N}{\partial V} + \sum_{t=1}^{N} \delta^l(u^l(t))\,(x^l(t))^{\mathrm{T}}, \tag{20}$$

$$\frac{\partial L_N}{\partial U} = \frac{\partial L_N}{\partial U} + \sum_{l=1}^{N} \delta^l(u^l(t))\,(h^l(t-1))^{\mathrm{T}}, \quad (21)$$

$$\frac{\partial L_N}{\partial a} = \frac{\partial L_N}{\partial a} + \sum_{l=1}^{N} \delta^l(v^l(t)), \quad (22)$$

$$\frac{\partial L_N}{\partial b} = \frac{\partial L_N}{\partial b} + \sum_{l=1}^{N} \delta^l(u^l(t)). \quad (23)$$

每次更新网络中的所有权值和偏置.

3 实验设计与结果分析

在实验中使用两层的 LSTM 递归神经网络和改进的 LSTM 神经网络,每层节点数为 1 024,该 LSTM 递归神经网络模型参数数量为 16.0 Million,在 TIMIT 数据集上进行实验. 输入为 36 维的 MFCC 原始声学特征. MFCC 特征提取中帧长取 256,窗函数使用汉明窗. 使用动量梯度下降学习函数,动量更新设置为 0.5,时延第一层设置为 -1,第二层设置为 -2. 在 TIMIT 数据集上将训练好的模型进行测试,如表 1 所示. 基于改进 LSTM 递归神经网络语音识别系统的错误率 *TER* 为 56.0%,与两层的 LSTM 递归神经网络相比较降低了 5.0%,表明改进的 LSTM 识别准确率显著高于标准 LSTM 的准确率.

表 1 基于改进 LSTM 递归神经网络语音识别系统

Tab. 1 A speech recognition system based on improved LSTM recurrent neural network

系统描述	*TER*/%
LSTM-RNN	61.0
改进 LSTM-RNN	56.0

笔者设计的实验中,对 LSTM 与改进的 LSTM 进行测试,比较不同迭代次数的 LSTM 和改进 LSTM 的准确率. 迭代次数设置为 1 ~ 10 次,随着迭代次数的增加,识别的准确率同时增加,改进的 LSTM 准确率高于 LSTM 的准确率. 比较不同迭代次数的 LSTM 和改进 LSTM 的准确率的实验结果如图 4 所示.

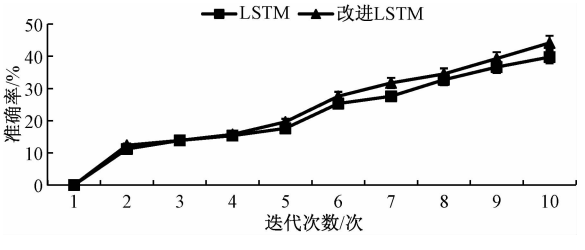

图 4 比较不同迭代次数的 LSTM 和改进 LSTM 的准确率

Fig. 4 Compare the LSTM of different iterations and the accuracy of the improved LSTM

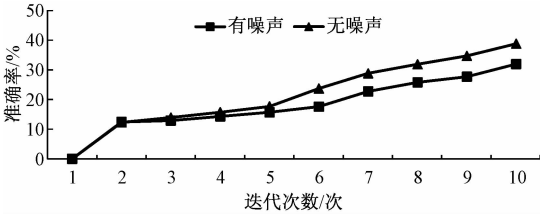

图 5 在有噪声和无噪声环境中对标准的 LSTM 进行测试的识别准确率

Fig. 5 identification accuracy of test for standard LSTM in a noisy and noiseless environment

在有噪声和无噪声环境中对标准的 LSTM 进行测试,随着迭代次数的增加,识别准确率上升,在无噪声环境中测试标准的 LSTM 的准确率普遍高于在有噪声环境中测试的准确率. 在有噪声和无噪声环境中对标准的 LSTM 进行测试的识别准确率如图 5 所示,结果表明改进的 LSTM 识别错误率较标准的 LSTM 识别错误率降低了 5%.

在有噪声和无噪声环境中对改进的 LSTM 进行测试,随着迭代次数的增加,识别的准确率上升,在无噪声环境中测试改进的 LSTM 的准确率普遍高于在有噪声环境中测试的准确率. 在有噪声和无噪声环境中对改进的 LSTM 进行测试的识别准确率如图 6 所示.

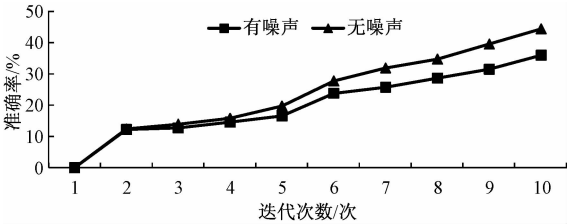

图 6 在有噪声和无噪声环境中对改进的 LSTM 进行测试的识别准确率

Fig. 6 identification accuracy of test for improved LSTM in a noisy and noiseless environment

4 结论

提出了基于改进的 LSTM 递归神经网络系统,增加从当前 CEC 到输出门的连接,更好地控制每一内存单元输出多少信息,将遗忘门和输入门合成了一个单一的更新门. 信息通过输入门和遗忘门将过去与现在的记忆进行合并,长短时记忆神经网络可以选择遗忘之前累积的信息,这样改进的 LSTM 模型可以学到长时期的历史信息,具有更强的鲁棒性. 采用基于改进的 LSTM 方法,在 TIMIT 数据集上进行实验. 基于改进的 LSTM 递归神经网络系统有效地提升了识别率,提高了语音识别系统的鲁棒性.

参考文献:

[1] LEI Y, SCHEFFER N, FERRER L. A novel scheme for speaker recognition using a phonetically-aware deep neural network[C] //Proc of IEEE International Conference on Acoustics, Speech and SignalProcessing. 2014:1695 – 1699.

[2] SAK H, SENIOR A, BEAUFAYS F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[J]. Computer Science, 2014 (11):338 – 342.

[3] SAKH, SENIOR A, BEAUFAYS F. Long short-term memory recurrent neural network architectures for largescale acoustic modeling[C] // Proc of Annual Conference of International Speech Communication Association. 2014:338 – 342.

[4] HINTON G, DENG L, YU. Deep neural networks for acoustic modeling in speech recognition[J]. IEEE signal processing magazine, 2012,29(6):82 – 97.

[5] 陶佰睿,郭琴,苗凤娟,等. 基于自适应 mel 滤波器组的 MFCC 特征提取的 SOC 设计[J]. 郑州大学学报(工学版),2016,37(3):11 – 15.

[6] 余凯,贾磊,陈雨强,等. 深度学习的昨天、今天和明天[J]. 计算机研究与发展,2013,50(9): 1799 – 1804.

[7] 史笑兴,顾明亮,王太君,等. 一种时间规整算法在神经网络语音识别中的应用[J]. 东南大学学报,1999,29(5): 47 – 51.

[8] CHEN K, YAN Z J, HUO Q. Training deep bidirectional LSTM acoustic model for LVCSR by a context sensitive-chunk BPTT approach[R]. In Proceedings of Interspeech,2015.

[9] CHEN X, WANG Y Q, LIU X Y, et al. Efficient gpu-based training of recurrent neural network language models using spliced sentence bunch[R]. Proceedings of Interspeech, 2014.

[10] PASCANU R, GULCEHRE C, CHO K, et al. How to construct deep recurrent neural networks[C]. Ar Xiv: 1312.6026, 2013.

[11] 孙志军,薛磊,许阳明,等. 深度学习研究综述[J]. 计算机应用研究,2012,29(8): 2806 – 2810.

[12] SCHMIDHUBER J. Deep learning in neural networks: An overview[J]. Neural networks, 2015(61):85 – 117.

[13] DU J, DAI L R, HUO Q. Synthesized stereo mapping via deep neural networks for noisy speech recognition [C] //2014 IEEE International Conference on Acoustics, Speech anold Signal Processing (ICASSP). Washington DC: IEEE,2014: 1764 – 1768.

[14] XU Y, MO T, FENG Q. Deep learning of feature representation with multiple instance learning for medical image analysis[C] // Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington DC: IEEE, 2014: 1626 – 1630.

[15] ABDEL-HAMID O, MOHAMED A, JIANG H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C] // In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012:4277 – 4280.

[16] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]. arXiv preprint arXiv:1409.0473, 2014.

Research on Speech Recognition Based on Improved LSTM Deep Neural Network

ZHAO Shufang, DONG Xiaoyu

(Institute of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: The language model based on neural network LSTM structure, the LSTM structure used in the hidden layer unit, the structure unit comprises a storage unit to store information for a long time, which has a good memory for historical information. But the LSTM in the current input information state does not affect the final output information of the output gate, get less historical information. To solve the above problems, this paper puts forward based on improved LSTM (long short-term memory) modeling method of network model. The model increases the connection from the current input gate to the output gate, and simultaneously combines the oblivious gate and the input gate into a single update gate. The door keeper input and forgotten past and present memory consolidation, can choose to forget before the accumulation of information, the improved LSTM model can learn the long history of information, solve the drawback of the LSTM method is more robust. This paper uses the neural network language LSTM model based on the improved model on TIMIT data sets show that the accuracy of test. The results illustrate that the improved LSTM identification error rate is 5% lower than the standard LSTM identification error rate.

Key words: long-short term memory(LSTM); deep neural network; speech recognition