基于残差的门控循环单元

张忠豪 [1] 董方敏 [1] 胡 枫 [1] 吴义熔 [1, 2] 孙水发 [1, 2]

摘 要 传统循环神经网络易发生梯度消失和网络退化问题. 利用非饱和激活函数可以有效克服梯度消失的性质, 同时借鉴卷积神经网络中的残差结构能够有效缓解网络退化的特性, 在门控循环神经网络 (Gated recurrent unit, GRU) 的基础上提出了基于残差的门控循环单元 (Residual-GRU, Re-GRU) 来缓解梯度消失和网络退化问题. Re-GRU 的改进主要包括两个方面: 1) 将原有 GRU 的候选隐状态的激活函数改为非饱和激活函数; 2) 在 GRU 的候选隐状态表示中引入残差信息. 对候选隐状态激活函数的改动不仅可以有效避免由饱和激活函数带来的梯度消失问题, 同时也能够更好地引入残差信息, 使网络对梯度变化更敏感,从而达到缓解网络退化的目的. 进行了图像识别、构建语言模型和语音识别 3 类不同的测试实验, 实验结果均表明, Re-GRU 拥有比对比方法更高的检测性能, 同时在运行速度方面优于 Highway-GRU 和长短期记忆单元. 其中, 在语言模型预测任务中的 Penn Treebank 数据集上取得了 23.88 的困惑度, 相比有记录的最低困惑度, 该方法的困惑度降低了一半.

关键词 深度学习, 循环神经网络, 门控循环单元, 残差连接

引用格式 张忠豪, 董方敏, 胡枫, 吴义熔, 孙水发. 基于残差的门控循环单元. 自动化学报, 2022, **48**(12): 3067−3074

DOI 10.16383/j.aas.c190591

Residual Based Gated Recurrent Unit

ZHANG Zhong-Hao[1] DONG Fang-Min[1] HU Feng[1]
WU Yi-Rong[1, 2] SUN Shui-Fa[1, 2]

Abstract Traditional recurrent neural networks are prone to the problems of vanishing gradient and degradation. Relying on the facts that non-saturated activation functions can effectively overcome the vanishing gradient problem, and the residual structure in convolution neural network can effectively alleviate the degradation problem, we propose a residual−gated recurrent unit (Re-GRU) which leverages gated recurrent unit (GRU) to alleviate the problems of vanishing gradient and degradation. There are two main improvements in Re-GRU. One is to replace the activation function of the candidate hidden state in GRU with the non-saturated activation function. The other is to introduce the residual information into the candidate hidden state representation of the GRU. The modification of candidate hidden state activation function can not only effectively avoid vanishing gradient caused by non-saturated activation function, but also introduce residual information to make the network more sensitive to gradient change, so as to alleviate the degradation problem. We conducted three kinds of test experiments, including image recognition, building language model, and speech recognition. The results indicate that our proposed Re-GRU has higher detection performance than other 6 methods. Specifically, we achieved a test-set perplexity of 23.88 on the Penn Treebank data set in language model prediction task, which is one half of the lowest value ever recorded.

Key words Deep learning, recurrent neural networks, gated recurrent unit, skip connect

Citation Zhang Zhong-Hao, Dong Fang-Min, Hu Feng, Wu Yi-Rong, Sun Shui-Fa. Residual based gated recurrent unit. *Acta Automatica Sinica*, 2022, **48**(12): 3067−3074

在过去的十几年里, 深度学习的提出对全球各个领域带来了巨大的影响. 深层神经网络、卷积神经网络和循环神经网络 (Recurrent neural network, RNN) 等神经网络模型被广泛应用于各个领域. 其中, 循环神经网络具有捕获长序依赖的能力, 因此被广泛应用于语音识别[1]、语言建模[2]、机器翻译[3] 等自然语言处理[1−4] 领域. 然而, 普通循环神经网络会因为梯度消失[5] 和梯度爆炸问题而变得不稳定, 于是学者们提出基于长短期记忆单元 (Long short-term memory, LSTM) 的时间递归神经网络[1, 6] 来缓解梯度消失和梯度爆炸问题. 虽然 LSTM 确实有效, 但其门限繁杂, 于是近些年有许多针对 LSTM的改良方案被提出, 其中门控循环单元 (Gated recurrent unit, GRU)[7] 是 LSTM 最具代表性的一种改进方案.

深度学习的成功主要归因于它的深层结构[8−9], 然而训练一个深层网络是较为困难的事. 随着网络层数的增加, 梯度消失、梯度爆炸、网络退化[10] 等问题会导致模型被损坏. 为了能够进行更深的网络训练, 目前已有多种深层前馈神经网络的结构被提出, 最具代表性的有高速公路网络[9], 用于卷积神经网络的残差网络[11], 以及最近被提出的能够进行更深网络训练的简单循环单元 (Simple recurrent units, SRU)[12−13].

在循环神经网络体系中, 因为通常使用了饱和激活函数, 所以很少会发生梯度爆炸问题, 但是由饱和激活函数而带来的梯度消失问题却很常见. 虽然 LSTM 和 GRU 相比传统的 RNN 是具备缓解梯度消失问题的能力, 但实际上这种缓解是有限的, 这个问题将在后文通过实验来具体展现. 在循环神经网络中也存在着网络退化问题, 导致循环神经网络的性能随着网络层数的增加而越来越糟糕. 采用高速公路网络的方法能够缓解网络的退化问题, 但是这种方法会增加网络参数量和训练耗时[11]. 近两年备受关注的 SRU 网络也包含了类似高速公路网络的结构[13], 同时 SRU 舍去了循环单元中的时间参数, 所以在运行快速的同时在一些任务中也能够进行更深的网络训练.

本文通过对 GRU 结构的深入研究, 发现通过修改其候选隐状态的激活函数并添加残差连接, 可以有效地解决原始 GRU 的梯度消失和网络退化问题. 而对于使用了非饱和激活函数而可能导致的梯度爆炸隐患, 本文则是采用了批标准化 (Batch normalization, BN)[14] 的方法来解决. 在本文的 3 类不同对比实验中, 本文设计的 (Residual-GRU, Re-GRU) 在 3 类实验中均取得了比 GRU、LSTM、Highway-

收稿日期 2019-08-18 录用日期 2020-01-17

Manuscript received August 18, 2019; accepted January 17, 2020

国家自然科学基金 (U1703261, 61871258), 国家重点研发计划 (2016-YFB0800403) 资助

Supported by National Natural Science Foundation of China (U17-03261, 61871258) and National Key Research and Development Project (2016YFB0800403)

本文责任编委 陈德旺

Recommended by Associate Editor CHEN De-Wang

1. 三峡大学计算机与信息学院 宜昌 443002 2. 智慧医疗宜昌市重点实验室 宜昌 443002

1. College of Computer and Information Technology, China Three Gorges University, Yichang 443002 2. Yichang Key Laboratory of Intelligent Medicine, Yichang 443002

GRU、SRU 等网络更好的效果, 并且在同样的配置下, 本文设计的 Re-GRU 比 LSTM 和 Highway-GRU 耗时更短.

1 现有技术

1.1 门控循环单元

门控循环单元 (GRU) 是 LSTM 的简化改进, GRU 和 LSTM 的共同点是每个神经元都是一个处理单元, 每个处理单元都包含了若干个门限, 门限可以判断输入的信息是否有用. 与 LSTM 不同的是, 每个 GRU 处理单元仅有两个门限, 并且 GRU 的单元只有一个时序输出, 所以 GRU 在保证能有效传递时序相关信息的条件下, 拥有更少的参数量. GRU 单元结构如图 1 所示, 公式定义如下:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$a_t = Tanh(W_a U_a (h_{t-1} r_t) + b_a) \tag{3}$$

$$h_t = (1 - z_t) h_{t-1} + z_t a_t \tag{4}$$

式中, x_t 表示当前层的 t 时刻输入值, h_{t-1} 是 $t-1$ 时刻的状态输出值, z_t 和 r_t 分别为 t 时刻的更新门和重置门, 更新门和重置门的激活函数 σ 是 Sigmoid 函数, a_t 为 t 时刻的候选隐状态, h_t 表示当前时间 t 的状态向量, $Tanh$ 是候选隐状态的双曲正切激活函数, 模型权重参数是 W_z、W_r、W_a、U_z、U_r 和 U_a, 偏置向量为 b_z、b_r 和 b_a.

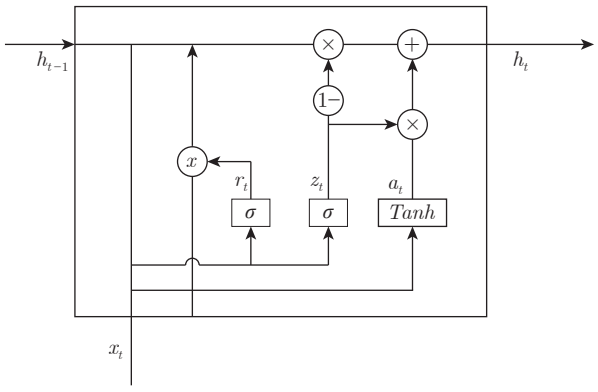


图 1 GRU 单元结构

Fig.1 GRU structure

1.2 高速公路网络

在普通神经网络中, 若网络层数过高且没有特殊限制, 往往会产生冗余的网络层, 冗余网络层存在不必要的计算, 这通常会导致网络性能变差. 要想让网络不产生冗余, 就需要对网络结构进行特殊的设计. 假设 x 是网络某层的输入, $F(x)$ 表示 x 经过了某种运算, $H(x)$ 是此层的输出, 则普通网络的输出为 $H(x) = F(x)$. 如果能在经过冗余的网络层时让输出 $H(x)$ 直接等同于输入 x, 也就是构成一个恒等映射 $H(x) = x$, 这样就能避免了信息通过不必要的计算, 从而保证了信息的有效传递, 实现消除网络退化的目的.

高速公路网络[9] 是一种能实现深度神经网络的结构, 高速公路网络通过一个辅助门控 T 来控制输出的信息. 假设某层的输入为 x, 原始的输出为 $F(x)$. 则当 $T = 1$ 时则新的输出 $H(x)$ 完全由 $F(x)$ 构成, 当 $T = 0$ 时则新输出 $H(x)$ 完全由上一层的信息 x 构成. 通过门控 T 即可让有效信息有效传递, 从而使网络层不冗余. 常用的高速公路网络结构见图 2, 公式定义如下:

$$T = \sigma(Wx + b) \tag{5}$$

$$C = 1 - T \tag{6}$$

$$H(x) = T \times F(x) + C \times x \tag{7}$$

式中, x 是输入, w 是权重, b 是偏置向量, $F(x)$ 是未经过高速公路网络之前的输出, C 与 T 是高速公路网络的门控, σ 是门控 T 的激活函数, $H(x)$ 是高速公路网络输出.

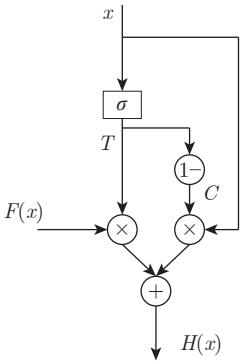


图 2 高速公路网络结构

Fig.2 Highway-network structure

高速公路网络通过门控的控制, 可以确保相有效信息能够畅通流动. SRU 网络也是运用了这种类似的结构[13], 使其网络能够更好地保证信息传输.

1.3 残差网络

残差网络[11] 的主要贡献是更好地解决了学习恒等映射函数问题, 并且没有增加网络参数数量. 残差网络将拟合恒等映射函数 $H(x) = x$ 转化为优化残差函数 $F(x) = H(x) - x$. 当残差函数中 $F(x) = 0$ 时就构成了 $H(x) = x$ 的恒等映射, 这样就避免了多余网络层产生的冗余, 使得网络在梯度下降的时候能够有效地应对网络退化问题[11]. 有研究表明, 网络退化的原因是权重矩阵的退化[10], 而残差连接的不对称性能够很好地应对权重矩阵退化问题.

残差网络与高速公路网络不同在于: 高速公路网络认为需要通过门限来控制残差的量, 而残差网络认为残差是随时存在的, 所以在计算的时候不需要门限的限制. 通过建立与之前层的信息通道, 使得网络可以堆积残差, 从而网络可以进行更深层的训练. 所以残差网络能更有效地学习残差映射且不会有更高的复杂度. 残差网络公式见式 (8), 原始残差网络如图 3 所示[11]. 而在没有残差计算的时候, 输出 y 则完全由 $\sigma(Wx^l + b^l)$ 提供.

$$y = \sigma((Wx^l + b^l) + x^{l-1}) \tag{8}$$

式中, x^l 表示 l 层的输入, x^{l-1} 表示 $l-1$ 层的输入, y 表示 l 层的输出, b^l 为 l 层的偏置项, σ 为激活函数.

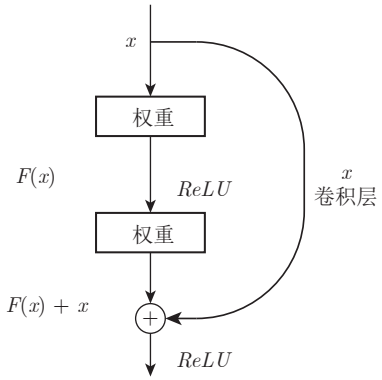


图 3 残差网络结构[11]

Fig. 3 Residual-Networks structure[11]

2 残差门控循环单元

2.1 残差门控循环单元结构

在循环神经网络中, 梯度消失和网络退化尤为严重. 针对这两个问题, 本文通过对 GRU 改进来解决. 在 GRU 的算法公式中最核心的公式为候选隐状态式 (3). 式 (3) 的输出值和前一时刻隐状态的输出值共同决定了 GRU 隐状态的最终输出. 所以本文的改进主要针对 GRU 的候选隐状态公式, 主要分为以下 3 点:

1) 非饱和激活函数

将 GRU 的候选隐状态的激活函数改为线性整流函数 (Rectified linear unit, ReLU), 这样能够让本文的改进网络能够很好地避免由饱和函数引起的梯度消失, 进而能够应对更深度的网络训练[15]. 常见的 ReLU 函数为:

$$ReLU(g(x)) = \begin{cases} f(x), & f(x) \geq 0 \\ 0, & f(x) < 0 \end{cases} \tag{9}$$

式中, $ReLU$ 函数的输入是 $f(x)$, $f(x)$ 表示 x 经过某种运算.

由于 $ReLU$ 在负半区的导数为 0, 所以一旦神经元激活后的值为负, 则梯度为 0. 当 $f(x)$ 大于 0 时, 有:

$$\frac{\partial ReLU(f(x))}{\partial x} = \frac{\partial f(x)}{\partial x} \tag{10}$$

$ReLU$ 激活函数可以保证信息传输更加直接, 相比饱和的激活函数, $ReLU$ 不存在饱和激活函数带来的梯度消失问题, 且能更好地配合残差信息的传递. 将式 (3) 改为:

$$a_t = ReLU(W_a x_t + U_a(h_{t-1} r_t) + b_a) \tag{11}$$

在过去的循环神经网络中, 使用无边界的非饱和激活函数通常会产生梯度爆炸问题. $ReLU$ 是非饱和激活函数的代表, 也存在梯度爆炸问题. 而将非饱和激活函数与批标准化技术相结合, 可以有效地使梯度爆炸问题得到有效的缓解[11, 16]. 其中文献 [16] 在 GRU 上进行了类似的改进, 也表明了在 GRU 网络上使用非饱和激活函数的有效性. 本文选择 ReLU 是由于其具有代表性, 使用其他类似的非饱和激活函数亦是可行的[17].

2) 添加残差连接

本文参考卷积神经网络中残差网络的方式来对 GRU 进行改进, 从而解决 GRU 中的梯度消失和网络退化问题. 具体地, 本文是将残差连接放在式 (11) 中. 对于引入的残差信息, 本文使用的是前一层的还未激活的候选隐状态值 net_a^{l-1}, 而不是前一层的隐状态输出值 h_{t-1}^{l-1}, 原因是未激活的值相比激活后的值具有更多的原始信息. 本文所设计的改进方案与卷积神经网络中的残差网络不同, Re-GRU 的每一层都有残差连接. 改进后的隐状态公式为:

$$a_t^l = ReLU(net_a^l) \tag{12}$$

$$net_a^l = (W_a^l x_t^l + U_a^l(h_{t-1}^l r_t^l) + b_a^l) + V^l net_a^{l-1} \tag{13}$$

式中, a_t^l 表示 l 层 t 时刻的候选隐状态的输出, net_a^{l-1} 为 $l-1$ 层的还未激活的候选隐状态值, h_{t-1}^l 表示 $t-1$ 时刻的 l 层状态向量, net_a^l 为 l 层的还未激活的候选隐状态, V^l 为第 l 层的维度匹配矩阵, 当网络上层和下层的维度相同时, 则不需要维度匹配矩阵.

此外, 也有研究者尝试过在循环神经网络上建立残差连接[13], 也设计过在 GRU 公式的其他公式中添加残差, 例如直接在式 (4) 中添加残差连接, 或在式 (3) 的激活函数之外添加残差连接, 而通过大量实验和算法推导发现本文所提方案是实验中性能最佳的改进方案.

3) 批标准化

批标准化[14] 通过对每个训练小批量的每个层的预激活的均值和方差进行规范化来解决数据内部协变量偏移, 同时也能够加速训练工程和提高系统性能. 同时也能够加速训练工程和提高系统性能. 使用批标准化可以缓解非饱和激活函数造成的梯度爆炸问题[16]. 本文通过改变 GRU 的激活函数同时添加残差连接, 再运用批标准化的优秀性质, 可以达到消除传统 GRU 中的梯度消失和网络退化的目的. 结合批标准化之后本文的 Re-GRU 第 l 层的单元结构见图 3, 结构单元公式如下:

$$z_t^l = \sigma(BN(W_z^l x_t^l) + U_z^l h_{t-1}^l) \tag{14}$$

$$r_t^l = \sigma(BN(W_r^l x_t^l) + U_r^l h_{t-1}^l) \tag{15}$$

$$a_t^l = ReLU(net_a^l) \tag{16}$$

$$net_a^l = BN(W_a^l x_t^l) + U_a^l(h_{t-1}^l r_t^l) + V^l net_a^{l-1} \tag{17}$$

$$h_t^l = (1 - z_r^l) h_{t-1}^l + z_r^l a_t^l) \tag{18}$$

式中, BN 表示采用的批标准化. 由于批标准化的性质就是消除偏差, 所以式 (14)、式 (15) 和式 (17) 中的偏置向量 b 被忽略.

2.2 残差门控循环单元的反向传播

由于循环神经网络具备层数空间, 同时也具备时序性, 所以在循环神经网络中存在反向传播和沿时间的反向传播. 本文设计的 Re-GRU 是基于神经网络上下层而建立的残差网络, 所以本文仅对 Re-GRU 的候选隐状态的反向传播展开讨论.

使用 $F(m)$ 对式 (17) 中的部分函数进行等效替换, 替换后的结构如图 4 所示.

$$F(m) = W_a x_t + U_a(h_{t-1} r_t) \tag{19}$$

假设此时时刻为 t, 设 L 为网络当前层数, 设 l 为要计算的误差项 $\delta_{a,t}^l$ 所在层数, L 层到 l 层之间有若干个网络层. 设网络每层的神经元个数相同, 此时式 (17) 的维度匹配矩阵 V 可以被忽略, 则式 (17) 可以简化为:

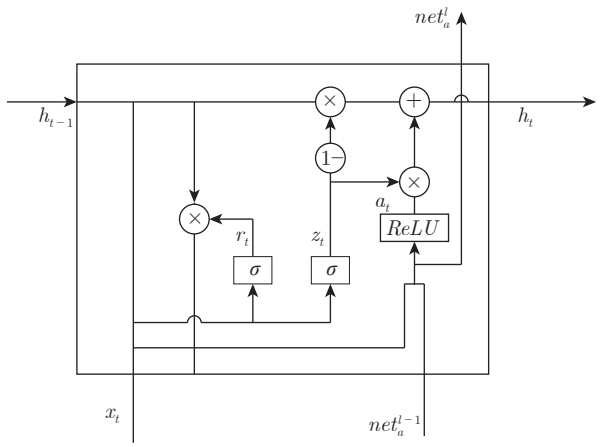


图 4 Re-GRU 结构

Fig. 4 Re-GRU structure

$$net_a^L = (W_a^L x_t^L + U_a^L (h_{t-1}^L r_t^L)) + net_a^{L-1} = F(m)^L + net_a^{L-1} = net_a^l + \sum_{i=l}^{L-1} F(m)^i \tag{20}$$

所以对 $net_{a,t}^L$ 求 $net_{a,t}^l$ 的偏导, 可展开为:

$$\frac{\partial net_a^L}{\partial net_a^l} = \frac{\partial \left(net_a^l + \sum_{i=l}^{L-1} F(m)^i \right)}{\partial net_a^l} = 1 + \frac{\partial \left(\sum_{i=1}^{L-1} F(m)^i \right)}{\partial net_a^l} \tag{21}$$

由于 ReLU 在负半区的导数为 0, 一旦神经元激活后的值为负, 则梯度为 0, 也就相当于不进行训练. 所以当 $net_{a,t}^L$ 大于 0 时, 有:

$$a_t^L = ReLU(net_a^L) = net_a^L \tag{22}$$

结合 ReLU 的性质, 在梯度大于 0 时有:

$$\frac{\partial a_t^L}{\partial net_a^l} = \frac{\partial (ReLU(net_a^L))}{\partial net_a^l} = \frac{\partial net_a^L}{\partial net_a^l} \tag{23}$$

所以根据链式求导法则, 候选隐状态的第 l 层误差项 $\delta_{a,t}^l$ 为:

$$\delta_{a,t}^l = \frac{\partial E_t}{\partial net_a^l} = \frac{\partial E_t}{\partial h_t^L} \frac{\partial h_t^L}{\partial a_t^L} \frac{\partial a_t^L}{\partial net_a^L} \frac{\partial net_a^L}{\partial net_a^l} = \frac{\partial E_t}{\partial h_t^L} \left(1 + \frac{\partial \left(\sum_{i=1}^{L-1} F(m)^i \right)}{\partial net_a^l} \right) z_t^L \tag{24}$$

式中, E 为误差, h_t^L 表示 L 层 t 时刻的状态向量, z_t^L 为 L 层更新门门限.

Re-GRU 的权重以及权重都使用来计算更新. 由式 (24) 可知, 在改变激活函数并添加残差连接后, Re-GRU 第 l 层候选隐状态误差项 $\delta_{a,t}^l$, 就能避免因 L 层到 l 层之间的多层连续相乘而导致的梯度消失. 通过 Re-GRU 的反向传播, 可以表明本文的 Re-GRU 相比传统的循环神经网络更有利于有效信息在层之间进行传递, 也表明了 Re-GRU 对梯度变化的敏感性, 达到了解决网络退化的目的.

3 实验和结果分析

3.1 数据集和任务

本文实验中的神经网络都是在 Linux 系统上利用 Pytorch 平台搭建神经网络完成模型训练, 都使用 NVIDIA GeForce GTX 1060 显卡进行加速训练. 本文分别对图像识别、语言模型预测和语音识别三类情况各进行了实验. 图像识别使用 MNIST数据集[18], 语言模型预测使用 PTB 数据集[2−3, 19] 和 WikiText-2 数据集[20], 语音识别使用 TIMIT 数据集[1, 21]. 为了方便本文更快完成地实验, 并且为了公平性, 本文实验中的每种循环神经网络都只使用单向的循环网络结构.

3.2 MNIST 数据集手写字识别任务

本文的第 1 个实验任务是图像识别, 采用了 MNIST 手写字数据集. 该数据集包含了 60 000 张 28 × 28 像素图片的训练集和 10 000 张 28 × 28 像素图片的测试集, 选择这个数据集主要是它能很好地展现梯度消失现象, 并且小数据集很方便进行模型试验. 在此数据集上, 本文使用了 RNN、GRU、LSTM、RNN-relu、GRU-relu、Highway-GRU、Re-GRU、SRU 来进行对比实验. 其中的 RNN-relu 是将激活函数替换成 ReLU 的 RNN, GRU-relu 是将候选隐状态激活函数换成 ReLU 的 GRU, Highway-GRU 是使用了高速公路网络结构的 GRU. 本文对每个神经网络都使用相同的配置: 均使用交叉熵损失函数作为损失函数; 梯度下降优化器均使用均方根传递 (Root mean square propagation, RMSProp); 均使用一层全连接层来匹配输出维度和目标维度; 隐藏层的神经元均为 64 个, 初始学习率均为 0.01. 所有网络分别进行 1、2、7、10、14 层的网络训练.

实验的评估标准为识别测试集的准确率. 具体实验结果见表 1. 由表 1 可以看出, RNN、GRU、LSTM 在网络层数增加时, 其模型的准确率则逐渐降低. 在 7 层之后 RNN、GRU、LSTM 模型的识别精度降低到约 10%. 由于这个任务要做的是识别一个手写字图到底是 0 ~ 9 中的哪一个数字, 而随机给出一个数字命中正确的概率就为 10%, 所以 10% 的准确率相当于随机的预测, 此时的模型已是无效模型.

本文对 RNN 和 GRU 的激活函数进行修改, 将饱和的激活函数 (RNN-relu) 改为非饱和激活函数 (GRU-relu). 由表 1 可以看出, RNN-relu 和 GRU-relu 相比, RNN 和 GRU 具有更好的表现. 这种效果的提升主要是使用了非饱和激活函数, 从而避免了饱和激活函数带来的梯度消失问题. 图 5 展示了同为 7 层的 GRU、GRU-relu、Re-GRU 的训练过程中的损失值变化曲线. 通过对比图 5 中的 GRU 和 GRU-relu 的损失值变化, 可以很直观地看出, 在使用饱和激活函数时, GRU 深层网络的梯度无法得到有效更新, 使得损失函数不能有效的降低, 而使用了非饱和激活函数, 则能避免这种因使用了饱和激活函数而导致的梯度消失问题.

但是, 简单地修改激活函数只能解决梯度消失问题, 却并不能应对网络退化问题. 从 2 ~ 14 层准确率变化情况可以看出, GRU-relu 和 RNN-relu 的准确率随着层数递增而不断下降, 到第 14 层时两者的模型都发生损坏, 这种情况主要是由于网络退化而导致的. Highway-GRU 的结果

表 1　MNIST 数据集测试结果 (%)
Table 1　MNIST dataset test results (%)

模型	1 层网络	3 层网络	5 层网络	7 层网络	9 层网络
RNN	52	12	9	10	10
GRU	92	92	11	11	10
LSTM	94	91	10	9	10
RNN-relu	67	72	63	56	9
GRU-relu	93	93	80	76	11
SRU	86	94	93	93	93
Highway-GRU	89	95	94	92	33
Re-GRU	97	96	94	95	94

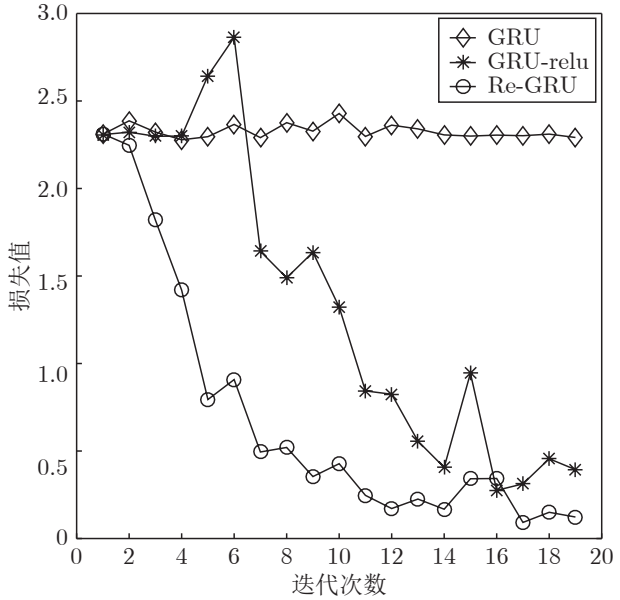


图 5　7 层网络的 GRU、GRU-relu、Re-GRU 在 MNIST 数据集上的损失变化曲线

Fig.5　Loss curve of GRU, GRU-relu and Re-GRU on the MNIST data set of a seven-layer network

可以表明利用不对称算法的方式是可以缓解网络退化的问题. 对于 Highway-GRU, 在 2 ~ 10 层时, 模型都表现较好, 但到了第 14 层时, 模型也发生了问题, 使得准确率大幅度下跌, 这种情况是 Highway-GRU 模型在网络层数较深时产生了梯度问题. 对于 SRU, 能维持较好模型效果的主要原因是其结构单元舍弃了时间信息参数, 所以在不需要时序信息的图像任务中有较好的表现.

通过对比表 1 可以看出, 本文设计的 Re-GRU相比其他模型不仅具备更好的识别效果, 同时在网络层数很深时依然是一个有效且良好的模型. 而在图 5 中, 对比 3 种模型的损失值变化曲线, 能够很清晰地反映出本文设计的 Re-GRU 损失值下降更加平滑且有效. 在更加深层的训练中, 本文设计的 Re-GRU 能够避免梯度问题, 依然能保持良好的准确率. 值得注意的是, 本文设计的 Re-GRU 在层数增加时, 准确率会小幅度降低. 这种情况主要是因为 MNIST 数据集较小, 很容易产生过拟合现象, 而 Re-GRU 并不能避免过拟合的发生.

3.3　PTB 和 WikiText-2 语言模型预测

本文第 2 个实验是语言模型预测, 使用的数据集是在语言模型预测中常用的数据集 Penn Treebank (PTB)[19]. 该数据集包含了 10000 个不同的词语和语句结束标记符, 以及标记稀有词语的特殊符号. 与图像任务不同, 语言模型任务是时序性的, 能更好地对比不同循环神经网络的性能. 本文使用了 Pytorch 官方的语言模型示例源码来完成 PTB 数据集的训练和测试, 其中的网络结构部分源码是本文重新构建的. 在此数据集上, 本文同样使用了 RNN、GRU、LSTM、RNN-relu、GRU-relu、SRU、Highway-GRU、Re-GRU 进行对比实验. 实验采用困惑度 (Perplexity, PPL) 评估标准, 一般来说 PPL 值越小则表明模型效果越好[2]. 实验中的每个循环神经网络配置都是完全相同的: 均设置了 650 个神经元; 嵌入层的大小均为 650, 均使用了批标准化, 丢弃率均为 50%. 每种网络分别进行 3、5、7 层的网络训练, 具体结果如表 2 所示, 其中逗号左为 PPL 值, 逗号右为迭代一次的运行时间.

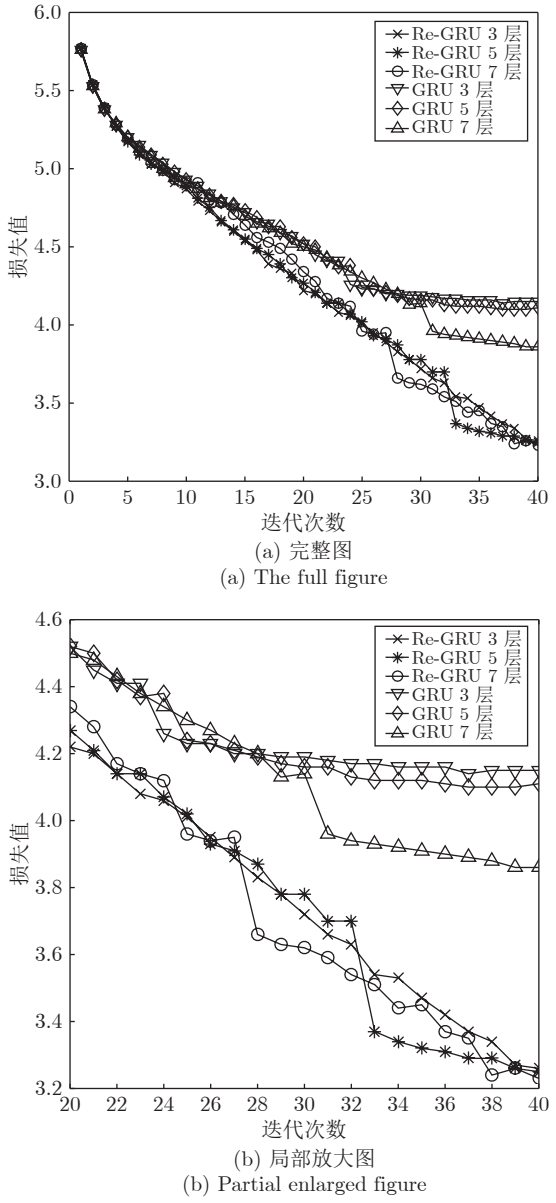
表 2　PTB 的测试结果 (PPL, s)
Table 2　PTB dataset test results (PPL, s)

模型	3 层网络	5 层网络	7 层网络
RNN	142.37, 149	135.56, 188	143.68, 214
GRU	59.73, 467	50.03, 584	50.25, 750
LSTM	56.42, 409	41.03, 542	84.27, 915
RNN-relu	125.83, 81	115.79, 164	117.32, 257
GRU-relu	96.71, 453	57.03, 602	90.14, 763
SRU	104.93, 206	124.18, 334	143.77, 432
Highway-GRU	99.77, 523	108.13, 834	88, 1176
Re-GRU	24.32, 378	23.88, 682	25.14, 866

表 2 中, 除了本文的 Re-GRU 之外, 表现最好的是 LSTM, 在第 5 层的时 PPL 达到了 41.03. 但是当网络层数升高到 7 层时, LSTM 的 PPL 值升高了一倍.

由表 2 可知, Highway-GRU 在实验 2 中表现不理想. 与 GRU 和 LSTM 的测试结果相对比, Highway-GRU 的表现要差很多. 虽然 Highway-GRU 能够在网络层数增加时有效降低 PPL 值, 但是其时间消耗是相对最多的. 本文同时对 Highway-GRU 进行了更深的网络训练, 在 9 层网络时, Highway-GRU 的 PPL 值为 90.86, 11 层时为 90.64, 这样的结果相比其他的模型效果是较差的, 因为 Highway-GRU 的复杂结构并不适用于这个任务. SRU 与高速公路网络具有类似的结构, 而 SRU 舍弃了时序参数. 所以 SRU 虽然运行速度很快, 但是效果却并不理想.

由表 2 可以看出, 本文设计的 Re-GRU 不仅有着最优的 PPL 值, 而且在网络层数升高的同时依然保持良好的效果. 将 GRU 和和 Re-GRU在训练过程中每次迭代得到的损失值进行统计, 并绘制了损失值变化曲线 (见图 6). 其中图 6(a) 为完整变化曲线, 图 6(b) 为局部放大图. 通过对比图 6 的损失变化曲线可以明显地看出, 本文设计的 Re-GRU 在迭代时损失下降更加平滑且更快. 这正是由于本文设计的 Re-GRU 网络对梯度变化敏感, 使得 Re-GRU 在网络的空间上的反向传播能够更有效的传递信息, 从而使



(a) 完整图
(a) The full figure

(b) 局部放大图
(b) Partial enlarged figure

图 6 GRU 与 Re-GRU 在 PTB 数据集上的损失值变化曲线

Fig.6 Loss curve of GRU and Re-GRU on PTB dataset

模型能更有效的完成学习任务. 表 2 直观地表明, 本文设计的 Re-GRU 相比 GRU 和 LSTM 的 PPL 值降低了超过 50%, 并且本文设计的 Re-GRU 在网络较深时依然能够拥有较好的模型效果. 在时间消耗的对比上, 本文设计的 Re-GRU 低于 LSTM和 Highway-GRU. 值得一提的是, 在 PTB 数据集上, 本文设计的 Re-GRU 所得到的困惑度达到 23.88, 这个测试结果比此前最佳记录的效果提升超过了一倍[2, 22].

为了进一步探索网络模型在语言模型任务上的有效性, 本文采用了一个与 PTB 数据集相类似的 WikiText-2 数据集[20] 作为本文的第 3 个实验. WikiText-2 相比 PTB 数据集规模更大一倍. 本文挑选了 RNN、GRU、LSTM、SRU、Re-GRU 网络来完成实验, 均使用 7 层的网络结构, 其他参数设定与本文实验 2 的参数设置相同. 实验结果见表 3, 其中逗号左为 PPL 值, 逗号右边为一次迭代的时间消耗.

表 3 WikiText-2 的测试结果 (PPL, s)

Table 3 WikiText-2 dataset test results (PPL, s)

模型	7 层网络
RNN	155.43, 235
GRU	43.87, 618
LSTM	29.00, 733
SRU	159.39, 514
Re-GRU	23.88, 644

表 3 中, 本文设计的 Re-GRU 依然拥有最低的 PPL, 并且运行时间低于 LSTM. 通过表 3 与表 2的数据对比 可以发现, 各个网络的在 PTB 数据集和 WikiText-2 数据集上的实验效果基本类似, 表明了本文改进方法的有效性.

另外, 经过多次实验和分析, 实验所用各个网络模型效果较好的主要原因是使用了批标准化算法. 当不使用标准化时, 以上网络结构实际性能普遍都达不到表 2 和表 3 的结果. 当使用了批标准化技术后, 各网络结构的性能皆有较大提升, 而各个模型的性能差异主要是网络结构的不同, 其中本文设计的 Re-GRU 结构在表 2 和表 3 中均有最佳困惑度的表现.

3.4 TIMIT 数据集语音识别任务

本文第 4 个实验是语音识别, 主要是语音识别声学模型训练, 采用经典的 TIMIT 语音识别数据集[23]. 该数据集一共包含 6300 个句子, 由来自美国 8 个主要方言地区的 630 个人每人说出给定的 10 个句子, 所有的句子都在音素级别上进行了手动分割好和标记[21]. 语音识别也是一个时序的任务, 通过这个任务, 进一步证明了本文设计的 Re-GRU 的可靠性和适用性. 该数据集的训练集和测试集分别占 90% 和 10%. 数据的特征是, 首先采用 Kaldi 工具箱[24] 提取的 39 维梅尔频率倒谱系数; 接着, 使用 Kaldi 工具箱来完成语音识别基线任务, 标签是通过对原始训练数据集执行强制对齐过程而得到的. 有关的更多细节参见 Kaldi 的标准 s5 配置; 然后, 在 Pytorch 平台上利用 Pytorch-Kaldi[21] 构建神经网络完成声学模型训练; 最后, 本文使用 Kaldi 的解码器完成语音识别效果评估任务, 评估标准采用标准的音素识别错误率 (Phone error rate, PER).

在语音识别实验中, 本文同样进行了 RNN、GRU、LSTM、RNN-relu、GRU-relu、SRU、Highway-GRU、Re-GRU 的循环神经网络模型训练, 并且补充了一个使用了非饱和激活函数的 GRU 改进网络 Light-GRU (Li-GRU)[21]. 所有神经网络模型分别进行了 3、5、7 层网络的实验, 最后使用 Softmax 分类器进行分类. 本文为每个循环神经网络都设置了每层 450 个神经元; 均设置了 20% 的遗忘率; 初始化方式均为正交初始化[25]; 批处理大小设置为每次 8 个句子; 损失函数均使用交叉熵损失函数; 优化器均使用 RMSProp 算法; 均使用了批标准化技术; 初始学习率均设置为 0.0008; 所有的模型进行 25 次迭代训练, 均设置了相同的学习率衰减. 具体实验结果见表 4, 其中逗号左为 PER 值, 逗号右为一次迭代的时间消耗.

由表 4 可以看出, RNN、GRU、SRU、Li-GRU 和 LSTM 在层数增加时错误率明显上升. Highway-GRU 相对于 GRU 有更低的音素识别错误率, 并且在层数增加时有着更好的模型效果, 但其消耗的时间却明显高于其他结构. 而本文设计的 Re-GRU 在 7 层网络时, 比 GRU 的音素识

表 4　TIMIT 的测试结果 (%, s)
Table 4　TIMIT dataset test results (%, s)

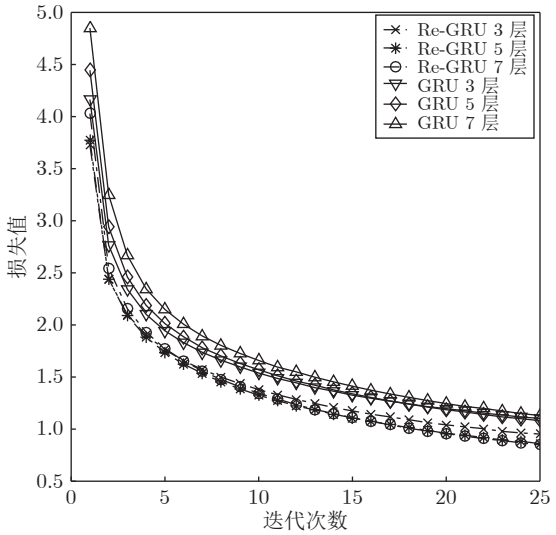
模型	3 层网络	5 层网络	7 层网络
RNN	22.5, 151	23.7, 225	23.9, 295
GRU	18.3, 389	18.2, 620	18.5, 854
LSTM	17.4, 478	17.2, 777	17.9, 1080
RNN-relu	18.3, 154	18.4, 239	18.6, 302
GRU-relu	17.3, 385	17.9, 616	17.8, 853
SRU	17.4, 404	18.3, 656	18.4, 924
Highway-GRU	18.0, 549	18.1, 908	17.5, 1294
Li-GRU	17.6, 287	17.9, 478	18.1, 630
Re-GRU	17.8, 427	17.5, 703	17.1, 984

别错误率低 1.4%, 比 LSTM 低 0.8%, 比 Li-GRU 低 1%. 此外, 在训练耗时方面, Re-GRU 的时间消耗低于 LSTM 和 Highway-GRU.

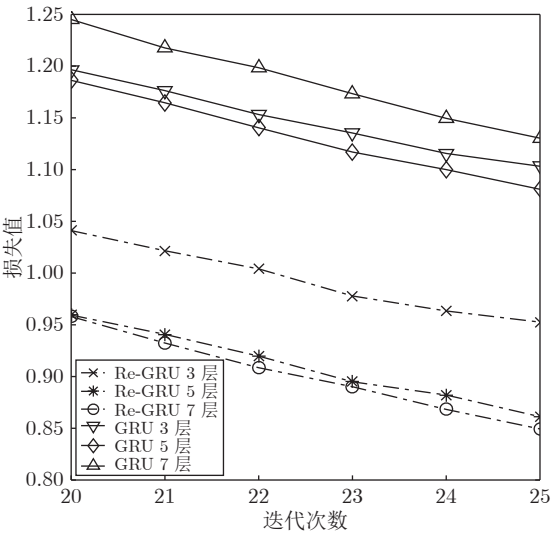
图 7(a) 为 GRU 和 Re-GRU 在训练过程中对训练集的损失值变化曲线, 为了让对比效果更明显, 本文选择了 3 层、5 层和 7 层网络并且没有使用学习率衰减. 其中图 7(b) 为图 7(a) 的第 20 ~ 25 次迭代的损失值曲线放大图. 由图 7 可以看出, 传统 GRU 的损失值随着网络层数的增加而越来越大, 而本文设计的 Re-GRU 能随层数增加而有效降低损失值. 损失值的有效降低使得本文设计的 Re-GRU 模型音素识别错误率能够低于其他模型.

当本文在 TIMIT 数据集的梅尔频率倒谱系数特征值上使用 Pytoch-Kaldi 平台构建 7 层网络的双向 Re-GRU 时, 模型的音素识别错误率低至 15.0%, 对比文献 [21] 方法, 本文设计的 Re-GRU 具有更优的模型效果.

此外, 本文采用了类似于文献 [11] 的构建跨越两层的残差连接, 将 Re-GRU 的跨越一层改为跨越两层. 跨越两层可以更远的传递信息, 通过这种方式可以使网络进行更远的残差学习. 在 TIMIT 数据集上进行额外的对比实验, 发现在同样网络深度下, 跨越两层的 Re-GRU 比跨越一层的 Re-GRU 音素识别错误率大约要更低 0.1% ~ 0.3%.



(a) 完整图
(a) The full figure



(b) 局部放大图
(b) Partial enlarged figure

图 7　GRU 与 Re-GRU 在 TIMIT 数据集上的损失值变化曲线

Fig.7　Loss curve of GRU and Re-GRU on TIMIT dataset

4　结束语

在循环神经网络体系中, 存在着梯度消失和网络退化问题, 本文基于 GRU 提出的 Re-GRU 具备解决梯度消失和网络退化问题的能力. 与传统循环神经网络相比, 本文的 Re-GRU 在网络层数较深时模型依然能有较好的性能, 并且本文的改进并没有增加网络的参数量. 相比传统的循环神经网络, 本文的 Re-GRU 有着更低的错误率和较低的训练耗时. 在缺点方面, 本文设计的 Re-GRU 不能够避免过拟合现象: 如果使用了特别深的网络, 虽然模型的损失值可能继续降低或者保持不变, 但模型效果却可能变差. 此外, 之前也对 RNN 和 LSTM进行了类似的改进并进行实验, 发现本文的改进方法在 RNN 上使用后能够相对 RNN 有较大效果提升, 但却并不适用于 LSTM. 通过理论分析和具体实验发现: 当将 LSTM 的两个或其中一个时序传输公式中的饱和激活函数修改为非饱和激活函数时, 都会导致模型发生梯度爆炸问题; 并且, 仅仅对 LSTM 直接添加残差连接未能取得较大效果提升.

References

1　Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic: 2013. 273−278

2　Mikolov T, Zweig G. Context dependent recurrent neural network language model. In: Proceedings of the 2012 IEEE Spoken Language Technology Workshop. Miami, USA: 2012. 234−239

3　Zhao B, Tam Y C. Bilingual recurrent neural networks for improved statistical machine translation. In: Proceedings of the 2014 IEEE Spoken Language Technology Workshop. South Lake Tahoe, USA: 2014. 66−70

4　Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for

natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445−1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, **42**(10): 1445−1465)

5 Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998, **6**(2): 107−116

6 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

7 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. Kuching, Malaysia, 2014: 3104−3112

8 Morgan N. Deep and wide: Multiple layers in automatic speech recognition. *Transactions on Audio, Speech, and Language Processing*, 2011, **20**(1): 7−13

9 Srivastava R K, Greff K, Schmidhuber J. Training very deep networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: 2015. 2377−2385

10 Orhan E, Pitkow X. Skip Connections eliminate singularities. In: Proceedings of the International Conference on Learning Representations. Vancouver, Canada: 2018. 1−22

11 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 770−778

12 Lei T, Zhang Y, Wang S I, Dai H, Artzi Y. Simple recurrent units for highly parallelizable recurrence. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: 2018. 4470−4481

13 Zhang Wen, Feng Yang, Liu Qun. Deep neural machine translation model based on simple recurrent units. *Journal of Chinese Information Processing*, 2018, **32**(10): 36−44
(张文, 冯洋, 刘群. 基于简单循环单元的深层神经网络机器翻译模型. 中文信息学报, 2018, **32**(10): 36−44)

14 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: 2015. 448−456

15 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: 2010: 249−256

16 Ravanelli M, Brakel P, Omologo M, Bengio Y. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2018, **2**(2): 92−102

17 Vydana H K, Vuppala A K. Investigative study of various activation functions for speech recognition. In: Proceedings of the 23th National Conference on Communications. Chennai, India: 2017. 1−5

18 Deng L. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012, **29**(6): 141−142

19 Marcus M P, Marcinkiewicz M A, Santorini B. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 2002, **19**(2): 313−330

20 Melis G, Dyer C, Blunsom P. On the state of the art of evaluation in neural language models. In: Proceedings of the 2018 International Conference on Learning Representations. Vancouver, Canada: 2018. 1−10

21 Ravanelli M, Parcollet T, Bengio Y. The Pytorch-Kaldi speech recognition toolkit. In: Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: 2019. 6465−6469

22 Oualil Y, Klakow D. A Neural Network approach for mixing language models. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA: IEEE, 2017. 5710−5714

23 Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 1990, **9**(4): 351−356

24 Arnab G, Gilles B, Luk′ as B, Ondrej G, Nagendra G, Mirko H, et al. The Kaldi speech recognition toolkit. In: Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding. Hawaii, USA: 2011. 59−64

25 Le Q V, Jaitly N, Hinton G E. A simple way to initialize recurrent networks of rectified linear units. arXiv preprint, 2015, arXiv: 1504.00941

张忠豪 三峡大学硕士研究生. 主要研究方向为人工智能和自然语言处理.
E-mail: zhangminecraftbiu@gmail.com
(**ZHANG Zhong-Hao** Master student at China Three Gorges University. His research interest covers artificial intelligence and nature language processing.)

董方敏 三峡大学教授. 主要研究方向为计算图形学, 计算机视觉和人工智能. E-mail: fmdong@ctgu.edu.cn
(**DONG Fang-Min** Professor at China Three Gorges University. His research interest covers computer graphics, computer vision and artificial intelligence.)

胡 枫 三峡大学硕士研究生. 主要研究方向为自然语言处理. E-mail: h18271692608@163.com
(**HU Feng** Master student at China Three Gorges University. His main research interest is nature language processing.)

吴义熔 三峡大学教授. 主要研究方向为人工智能和自然语言处理. E-mail: yirongwu@gmail.com
(**WU Yi-Rong** Professor at China Three Gorges University. His research interest covers artificial intelligence and nature language processing.)

孙水发 三峡大学教授. 主要研究方向为多媒体信息处理和智能信息处理. 本文通信作者.
E-mail: watersun@ctgu.edu.cn
(**SUN Shui-Fa** Professor at China Three Gorges University. His research interest covers multi-media information processing and intelligent information processing. Corresponding author of this paper.)