双向长短时记忆模型训练中的空间平滑正则化方法研究

李文洁[①②] 葛凤培[①②] 张鹏远*[①②] 颜永红[①②③]

[①](中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190)

[②](中国科学院大学 北京 100049)

[③](中国科学院新疆理化技术研究所新疆民族语音语言信息处理实验室 乌鲁木齐 830011)

摘 要：双向长短时记忆模型(BLSTM)由于其强大的时间序列建模能力，以及良好的训练稳定性，已经成为语音识别领域主流的声学模型结构。但是该模型结构拥有更大计算量以及参数数量，因此在神经网络训练的过程当中很容易过拟合，进而无法获得理想的识别效果。在实际应用中，通常会使用一些技巧来缓解过拟合问题，例如在待优化的目标函数中加入L2正则项就是常用的方法之一。该文提出一种空间平滑的方法，把BLSTM模型激活值的向量重组成一个2维图，通过滤波变换得到它的空间信息，并将平滑该空间信息作为辅助优化目标，与传统的损失函数一起，作为优化神经网络参数的学习准则。实验表明，在电话交谈语音识别任务上，这种方法相比于基线模型取得了相对4%的词错误率(WER)下降。进一步探索了L2范数正则技术和空间平滑方法的互补性，实验结果表明，同时应用这2种算法，能够取得相对8.6%的WER下降。

关键词：语音信号处理；空间平滑；双向长短时记忆模型(LSTM)；正则化；过拟合

中图分类号：TN912.34 **文献标识码**：A **文章编号**：1009-5896(2019)03-0544-07

DOI: 10.11999/JEIT180314

Spatial Smoothing Regularization for Bi-direction Long Short-term Memory Model

LI Wenjie[①②] GE Fengpei[①②] ZHANG Pengyuan[①②] YAN Yonghong[①②③]

[①](*Key Laboratory of Speech Acoustics and Content Understanding*, *Institute of Acoustics*, *Chinese Acadamy of Sciences*, *Beijing* 100190, *China*)

[②](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

[③](*Xinjiang Laboratory of Minority Speech and Language Information Processing*, *Xinjiang Technical Institute of Physics and Chemistry*, *Chinese Academy of Sciences*, *Urumqi* 830011, *China*)

Abstract: Bi-direction Long Short-Term Memory (BLSTM) model is widely used in large scale acoustic modeling recently. It is superior to many other neural networks on performance and stability. The reason may be that the BLSTM model gets complicated structure and computation with cell and gates, taking more context and time dependence into account during training. However, one of the biggest problem of BLSTM is overfitting, there are some common ways to get over it, for example, multitask learning, L2 model regularization. A method of spatial smoothing is proposed on BLSTM model to relieve the overfitting problem. First, the activations on the hidden layer are reorganized to a 2-D grid, then a filter transform is used to induce smoothness over the grid, finally adding the smooth information to the objective function, to train a BLSTM network. Experiment results show that the proposed spatial smoothing way achieves 4% relative reduction on Word Error Ratio (WER), when adding the L2 norm to model, which can lower the relative WER by 8.6% jointly.

收稿日期：2018-04-03；改回日期：2018-11-22；网络出版：2018-12-03

*通信作者： 张鹏远 pzhang@hccl.ioa.ac.cn

基金项目：国家重点研发计划重点专项(2016YFB0801203, 2016YFB0801200)，国家自然科学基金(11590770-4, U1536117, 11504406, 11461141004)，新疆维吾尔自治区科技重大专项(2016A03007-1)

Foundation Items: The National Key Research and Development Plan (2016YFB0801203, 2016YFB0801200), The National Natural Science Foundation of China (11590770-4, U1536117, 11504406, 11461141004), The Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (2016A03007-1)

Key words: Speech signal processing; Spatial smoothing; Long Short-Term Memory (LSTM); Regularization; Overfitting

1　引言

语音识别是将语音转成对应的文字的技术，语音识别系统主要包括声学模型、语言模型和解码器几个模块。声学模型要解决的问题是语音特征和建模单位之间的匹配性和区分性问题；语言模型从高层语义的角度为解码器提供进一步细化搜索路径的信息，从而提高解码结果的准确率。声学模型作为将语音特征和建模单元联系起来的基础，在语音识别中扮演着重要的角色。如何提高模型的鲁棒性和区分性，解决目前声学模型存在的局限，一直是声学模型建模的主要研究问题。目前语音识别系统中常用的声学模型大都是以神经网络(Neural Network, NN)隐马尔可夫(Hidden Markov Model, HMM)模型为基本框架，用HMM来建模音素帧间的时序关系，神经网络模型描述不同HMM状态的概率分布情况，常用的一些神经网络模型有前馈深度神经网络(Feedforward Deep Neural Network, FDNN)，递归神经网络(Recurrent Neural Network, RNN)，卷积神经网络(Convolution Neural Network, CNN)。由于语音是一种在时间维度上有着复杂变化性和相关性的信号，RNN可以建模这种序列信号的变化，所以在语音识别中，使用RNN作为声学模型可以取得比传统的前馈深度神经网络更好的效果。双向长短时记忆模型(Bidirectional Long Short-Term Memory, BLSTM)[1]是一种特殊的RNN，近年来，人们发现BLSTM网络可以更准确地建模语音的短时变化和较长时间跨度上的依赖关系，将其运用于语音识别领域可以大幅提高识别准确率[2]。

神经网络的训练准则是网络训练的目标，通常用到的是交叉熵(Cross Entropy, CE)准则，该准则是定义在帧级别的优化准则。由于语音信号是一个时序信号，所以更为匹配的优化准则应该是定义在整个序列上。在语音识别中，序列化准则一般为最大化正确的词序列或音素序列的概率，常见的序列化准则有最大互信息(Maximum Mutual Information, MMI)准则[3]，增强MMI(Boosted MMI, BMMI)[4]，最小音素错误[5]，最小贝叶斯风险[6,7]等。此外还可以增加状态聚类方法来提升性能[8]。将序列化准则应用于声学模型神经网络训练中，能够较大幅度降低语音识别的词错误率。但是在实际应用过程中，序列化准则下的神经网络模型面临着过拟合的问题。网络训练在过拟合的情况下，训练集的损失在减小，但是测试集的损失不再变化或者开始增大，这样的网络能很地拟合训练数据，但是对于训练集中没有出现过的测试数据，泛化能力不够。针对这一问题，常用的方法是丢弃法(dropout)[9]或者在网络中加入一些正则化技术，比如L1正则，L2正则。很多正则方法的基本思想是限制网络容量[10]，比如在目标函数中加上某些参数的平方和，来控制参数优化过程中一些大的权重。

针对序列化准则下的BLSTM网络训练的过拟合问题，本文提出一种空间平滑的方法，可以理解为一种对网络激活值进行正则的方法。区别于一般的正则方法会将参数相对独立的进行处理，空间平滑方法先将BLSTM网络的激活值重组拼接成一个2维的网格，对网格进行滤波，得到的结果去匹配一些目标模式。将这个正则目标和序列化准则一起，进行模型参数优化。此外，本文还分析了空间平滑算法与L2正则方法的融合效果。实验表明，空间平滑算法可以有效地改善模型的泛化能力，显著提升语音识别性能；而且该方法与传统的L2正则方法具有较好的互补能力，融合后的模型可以进一步降低识别错误率。

本文的主要内容包含5节，第2节简要介绍了前期相关工作，L2正则化；第3节详细介绍本文提出的空间平滑算法原理及其在BLSTM深度神经网络模型上的具体实现；第4节是实验设计，结果分析和讨论；最后1节总结全文。

2　相关工作

2.1　BLSTM网络结构

在语音识别任务中，一般情况下BLSTM网络是一种性能优于DNN(Deep Neural Network)，TDNN(Time delay Deep Neural Network)，浅层CNN等模型的网络结构。BLSTM网络中存在一个特殊的结构叫记忆单元，记忆单元包括可以储存信息的细胞状态(cell state)，还有控制信息流的3个门：输入门(input gate)控制输入到细胞的信息流，输出门(output gate)控制从细胞输出的信息流，忘记门(forget gate)对细胞内部状态进行处理，可以自适应地调节或者忘记细胞的状态[2]。语音识别中常用的LSTM网络结构如图1，运算方法见式(1)—式(8)。其中i, f, o和c分别代表输入门、忘记门、输出门和细胞状态值，$\boldsymbol{W}$项是权重矩阵(比如$\boldsymbol{W}_{\mathrm{ix}}$是输入到输入门的权重矩阵)，$\boldsymbol{b}$代表偏

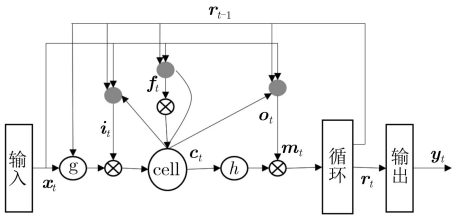

图 1 LSTM网络的记忆单元

置向量(比如$\boldsymbol{b}_i$就是输入门的偏置向量)，σ是sigmoid函数，$\odot$代表元素之间点乘，g和h是激活函数，在这里都使用tanh函数。如图1所示，$\boldsymbol{c}_t$代表当前时刻的LSTM单元中储存的信息，输入门$\boldsymbol{i}_t$控制流向细胞状态$\boldsymbol{c}_t$的输入，输出门$\boldsymbol{o}_t$控制着从细胞状态输出到网络其他部分信息。在将细胞值输出给循环连接之前，用忘记门$\boldsymbol{f}_t$对细胞状态的值进行加权，这样可以自适应地忘掉一些信息。

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{\mathrm{ix}}\boldsymbol{x}_t+\boldsymbol{W}_{\mathrm{im}}\boldsymbol{r}_{t-1}+\boldsymbol{W}_{\mathrm{ic}}\boldsymbol{c}_{t-1}+\boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\mathrm{fx}}\boldsymbol{x}_t+\boldsymbol{W}_{\mathrm{fm}}\boldsymbol{r}_{t-1}+\boldsymbol{W}_{\mathrm{fc}}\boldsymbol{c}_{t-1}+\boldsymbol{b}_f) \tag{2}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot g(\boldsymbol{W}_{\mathrm{cx}}\boldsymbol{x}_t+\boldsymbol{W}_{\mathrm{cm}}\boldsymbol{r}_{t-1}+\boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{o}_t=(\boldsymbol{W}_{\mathrm{ox}}\boldsymbol{x}_t+\boldsymbol{W}_{\mathrm{om}}\boldsymbol{r}_{t-1}+\boldsymbol{W}_{\mathrm{oc}}\boldsymbol{c}_t+\boldsymbol{b}_o) \tag{4}$$

$$\boldsymbol{m}_t=\boldsymbol{o}_t\odot h(\boldsymbol{c}_t) \tag{5}$$

$$\boldsymbol{p}_t=\boldsymbol{W}_{\mathrm{pm}}\boldsymbol{m}_t \tag{6}$$

$$\boldsymbol{r}_t=\boldsymbol{W}_{\mathrm{rm}}\boldsymbol{m}_t \tag{7}$$

$$\boldsymbol{y}_t=(\boldsymbol{p}_t,\boldsymbol{r}_t) \tag{8}$$

2.2 交叉熵(Cross Entropy, CE)准则和最大互信息(Maximum Mutual Information, MMI)[11]序列化准则

CE准则[12]是神经网络多分类问题常用的准则，目标函数为

$$F_{\mathrm{CE}}=-\sum_{u=1}^{U}\sum_{t=1}^{\mathrm{Tu}}\lg y_{\mathrm{ut}}(s_{\mathrm{ut}}) \tag{9}$$

其中，u代表句子，t表示时间，s_{ut}是句子u在t时刻的参考状态，$y(s)$是网络输出的预测值，$y_{\mathrm{ut}}(s_{\mathrm{ut}})$是网络在$s_{\mathrm{ut}}$状态上对应的输出概率。所以等式右边计算的是参考状态的分布和网络预测分布的交叉熵。

MMI准则的目标函数为

$$F_{\mathrm{MMI}}=\sum_{u}\lg\frac{p(\boldsymbol{O}_u|\boldsymbol{S}_u)^k P(\boldsymbol{W}_u)}{\sum_{w}p(\boldsymbol{O}_u|\boldsymbol{S})^k P(\boldsymbol{W})} \tag{10}$$

其中，u代表句子，$\boldsymbol{O}_u$是对于句子u观测序列，$\boldsymbol{S}_u$是句子u对应的状态序列，$\boldsymbol{W}_u$代表句子u在参考标注中的词序列，k是声学模型的影响系数，MMI准则是要最大化观测序列$\boldsymbol{O}_u=\{o_{\mathrm{u1}},o_{\mathrm{u2}},\cdots,o_{\mathrm{uTu}}\}$和参考序列$\boldsymbol{W}_u$的互信息[13]。

通常情况下，要进行MMI序列化训练，需要先训练一个CE的模型，并解出训练数据的解码网格(lattice)，其中包含了可能出现的一些观测序列，作为分母网格(denominator lattice)，然后在CE模型的基础上进行序列化，计算解码网格是一个很耗时的过程。2016年，Povey等人[11]提出了LF-MMI序列化训练，这种方法不需要预训练CE网络，也不需要提前解出信息网格，网络的训练时间比CE模型少，同时能够比CE准则获得相对约11.5%的词错误率下降。所以文献[11]的基线模型采用LF-MMI准则下的BLSTM网络，同时该模型还将CE准则乘以一定的权值作为学习的另一目标。

2.3 神经网络的L2范数正则

深度神经网络在训练过程中容易遇到过拟合的问题，通常使用网络正则化来减轻过拟合。常用的L2参数正则一般被认为是一种权值衰减的方法，他会在优化目标函数的同时，对网络中较大的权值进行衰减。实现方法是在目标函数上加上一个正则项，$\Omega(\theta)=(1/2)||\boldsymbol{w}||_2^2$，正则之后的模型的目标函数由原来的$J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})$变成$\tilde{J}(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})$

$$\tilde{J}(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})=\Omega(\theta)+J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y}) \tag{11}$$

文中用到的是L2参数正则的变体，$\Omega(\theta)=-(1/2)c||\boldsymbol{y}||_2^2$，用LF-MMI网络的输出代替网络权重来求取正则项，对网络输出向量$\boldsymbol{y}$求平方和，乘以系数之后加到目标函数中，得到最终的目标函数见式(12)

$$\tilde{J}(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})=-(1/2)c||\boldsymbol{y}||_2^2+J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y}) \tag{12}$$

3 本文提出的方法

3.1 空间平滑算法

神经网络激活值正则化处理，是对隐层激活值做一些正则变换。在训练过程中通过一些方法调节网络节点的激活值，使得它们的分布能够符合一些特定的模式。但是由于神经网络某一隐层节点是任意排序的，节点之间是相对独立的，很难相互关联，这可能会导致激活值的正则效果不理想。如果可以让这些激活值有一定的排序和关联，再让他们去匹配一些特定的目标模式，比如增加或减小激活值之间的差异性，可能会有利于神经网络正则。文献[14]也提出过类似的想法，文章提到平滑激活值的空间信息，模型正则可以取得更好的效果。

本文提出的方法引入了一种空间的排序，将BLSTM网络某层激活值的1维向量$\boldsymbol{h}_t^{(l)}$，组成一个2维的网格$\tilde{\boldsymbol{H}}_t^{(l)}$，比如网络某层的节点数是1024，可以当成一个32 × 32的网格。图2描述了如何将一

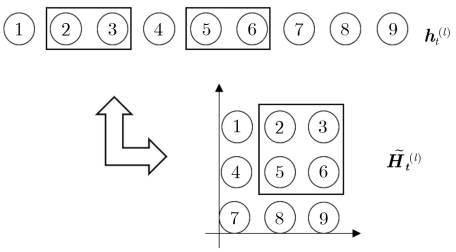

图 2 将激活值的1维向量拼成2维网格

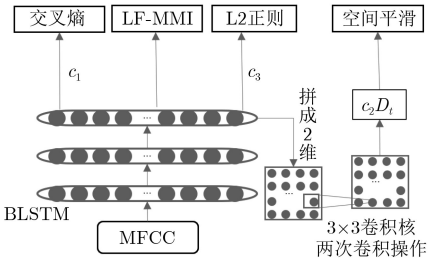

图 3 模型结构图

个节点数为9的隐层$\boldsymbol{h}_t^{(l)}$，拼成一个3×3的2维网格$\tilde{\boldsymbol{H}}_t^{(l)}$。

得到了激活值网格之后，可以对它进行一些变换，比如归一化，或者高通滤波。本文使用高通滤波器在激活值网格中引入空间平滑，滤波操作可以通过一个卷积操作来实现。用一个高通的卷积核，提取网格中相邻区域内节点的空间分布和关联的信息，经过两次这样的滤波之后，得到输出网格$\tilde{\boldsymbol{H}}^{*(l)}_t$，见式(13)。其中$k$是一个大小设定为$3\times3$的卷积核，权重分布中心为1，周围8个值都取–1/8，来达到对激活值进行平滑的效果。对于卷积核大小的选择，预先进行了一些实验，结果表明使用过大的卷积核，平滑正则的效果不理想，3×3是一个较为合适的取值。

$$\tilde{\boldsymbol{H}}^{*(l)}_t = \tilde{\boldsymbol{H}}^{(l)}_t * k * k \tag{13}$$

式中“*”表示卷积运算，经过这样的高通滤波器，可以使得临近节点有一定关联，如果训练目标是使得输出网格$\tilde{\boldsymbol{H}}^{*(l)}_t$趋于零，那么训练过程就会鼓励邻域的节点值更接近，所以网格上临近的网络节点就趋向于处在一个平滑的激活状态。

很多正则的方法都是先对网络的部分激活值进行变换[15]，然后使得变换之后的值能够匹配一定的目标模式，它可以是随时间变化的，也可以是时不变的，这些目标模式能够代表一定的网络的训练趋向。比如在L2参数正则中，就是先对网络的权重进行2范数变换，变换之后的值去匹配一个零的目标模式，这样网络训练的目标就会倾向于对网络的权值进行一定程度的衰减。类似地，文中采用的目标模式，为网格$\boldsymbol{G}_t^{(l)}=0$，计算变换之后的输出网格和目标模式之间的差异，$D_t=||\tilde{\boldsymbol{H}}^{*(l)}_t-\boldsymbol{G}_t^{(l)}||_2^2$，以减小这个差异为训练目标，进行神经网络训练。这样的训练目标，会让相邻区域内的节点趋向于处在一个相对平滑的激活状态。图3是网络结构图，包括多个训练目标，目标1是LF-MMI序列化准则，目标2是带权重的交叉熵(CE)准则，然后将本文提出的空间平滑方法作为神经网络目标函数的一个正则项。加入空间平滑之后的目标函数见式(14)，其中c_1, c_2分别表示交叉熵任务和空间平滑对目标函数的贡献。c_1采用Kaldi[16]中推荐的配置，取值为0.025；c_2的取值需要根据实验结果进行调节。最后还增加了在网络中使用L2正则的实验，用来对比观察2种正则方法的效果和叠加性，L2正则的计算方法见2.3节。

$$\begin{aligned}F_{obj}=&\sum_u \lg\frac{p(\boldsymbol{O}_u|\boldsymbol{S}_u)^k P(\boldsymbol{W}_u)}{\sum_W p(\boldsymbol{O}_u|\boldsymbol{S})^k P(\boldsymbol{W})}\\&-c_1\sum_{u=1}^{U}\sum_{t=1}^{\mathrm{Tu}}\lg y_{\mathrm{ut}}(s_{\mathrm{ut}})\\&-c_2||\tilde{\boldsymbol{H}}^{*(l)}_t-\boldsymbol{G}_t^{(l)}||_2^2-c_3L_2\mathrm{norm}\end{aligned}\tag{14}$$

3.2 BLSTM中的空间平滑

空间平滑任务的权重c_2是一个重要的参数，需要通过实验来调节，本文首先组织了一组实验，来对比不同权重对实验结果的影响。

考虑到BLSTM网络结构的特殊性，还需要详细探究应该对网络的哪个部分进行空间平滑，也就是需要调整组成输入网格的值，来对比评估空间平滑方法对BLSTM的影响。本文设计了以下几个对比实验，首先考虑将BLSTM 3个门(gate)的激活值分别来生成网格，然后对网络的细胞(cell)激活值进行处理，还可以对BLSTM的输出和循环部分进行空间平滑。实验采用的主要平滑的位置如下：

(1)位置1(p1)个门(gate)：$\tilde{\boldsymbol{H}}^{(l)}_{\mathrm{it}}=\mathrm{grid}\{\boldsymbol{i}_t^{(l)}\}$, $\tilde{\boldsymbol{H}}^{(l)}_{\mathrm{ot}}=\mathrm{grid}\{\boldsymbol{o}_t^{(l)}\}$, $\tilde{\boldsymbol{H}}^{(l)}_{\mathrm{ft}}=\mathrm{grid}\{\boldsymbol{f}_t^{(l)}\}$；

(2)位置2(p2)细胞值：$\tilde{\boldsymbol{H}}^{(l)}_t=\mathrm{grid}\{\boldsymbol{c}_t^{(l)}\}$；

(3)位置3(p3)网络输出值：$\tilde{\boldsymbol{H}}^{(l)}_t=\mathrm{grid}\{\boldsymbol{y}_t^{(l)}\}$；

(4)位置4(p4)循环值：$\tilde{\boldsymbol{H}}^{(l)}_t=\mathrm{grid}\{\boldsymbol{r}_t^{(l)}\}$。

4 实验及结果

4.1 实验设置

本文的实验使用Kaldi语音识别工具包，在Switchboard(Swbd)数据集上进行。Swbd数据是语音识别常用的数据集，训练数据中包含大约300 h

的英语电话对话交谈数据。测试集采用Hub5'2000数据中的一部分，eval2000数据集，eval2000数据时长约为3.8 h，由两部分构成，一部分是Swbd数据，这部分数据风格与训练集相近；另一部分是CallHome(CallHm)数据，这部分是和家人的电话交谈数据，风格更随意。本文分别对这两个子集进行测试，并且取两者的WER平均结果作为不同算法之间性能比较的最终评价指标，表格中用总计表示。文中实验的对比主要关注总计的结果。

实验先用13维MFCC和它的1, 2阶差分训练GMM-HMM模型，此外特征还做了说话人自适应(Speaker Adaptive Training, SAT)和特征空间最大似然线性回归(Feature-space Maximum Likelihood Linear Regression, FMLLR)变换，用这个GMM模型进行数据对齐。然后使用40维高分辨率MFCC训练神经网络模型，同时在输入特征中加入了100维的说话人(iVector)特征。

本文所用声学模型是3层的双向LSTM(BLSTM)模型，每层的前向和后向都有1024个节点，循环和非循环的节点都是256，网络输出维度是6050。网络每层的延迟是固定的，都是–3(+3)，这样将帧率降为原来的1/3，减少了计算量，可以加快训练速度。训练神经网络时，对数据进行速度扰动，把数据扩增为原来的3倍[16]。训练语言模型的语料是公开数据集Fisher数据和Swbd训练数据的所有标注文本。

4.2 实验结果

4.2.1 调节空间平滑的位置

本小结对3.2节中提到的4个不同的BLSTM位置的激活值进行平滑实验，模型使用LF-MMI准则，没有加入L2正则。进行这组实验之前，为了确定平滑权重应该选取的大致范围，本文预先试验了几组权重，结果表明平滑权重大于0.01时，识别词错误率会上升，权重取值在0.001附近时，词错误率会得到一定程度的下降。

BLSTM模型在不同位置通过微调空间平滑权重参数得到的实验结果见表1。结果表明，对BLSTM结构中细胞状态的激活值进行空间平滑，能够获得绝对0.4%的词错误率下降。在3个门(gate)采用空间平滑方法，也能一定程度改善识别结果，但是效果不如加在细胞值上明显，对此本文提出的猜想是，由于BLSTM结构3个门的作用是控制网络单元内的信息流动，修改它们激活值的分布对改善网络性能的帮助不大。分析表1中最后3行可以发现，对循环节点的值进行空间平滑处理时，网络的识别性能有所下降，文献[17]也提到过，在循环节点处使用正则的方法容易降低网络的性能。

表 1 不同位置空间平滑的结果

空间平滑位置	空间平滑权重(c)	CallHm WER (%)	Swbd WER (%)	总计WER (%)
无	无	20.0	10.3	15.2
P1	0.0020	19.9	10.4	15.2
P1	0.0010	19.9	10.0	15.0
P1	0.0007	20.0	10.3	15.2
P2	**0.0020**	**19.7**	**10.0**	**14.9**
P2	**0.0010**	**19.7**	**9.8**	**14.8**
P2	0.0007	19.9	9.8	15.0
P3	0.0020	20.1	10.3	15.2
P3	0.0010	20.0	9.8	15.0
P3	0.0007	20.0	10.1	15.1
P4	0.0010	20.9	10.6	15.8
P4	0.0007	20.6	10.3	15.5
P4	0.0006	20.5	10.6	15.6

所以接下来的实验都选择在BLSTM结构的细胞状态(cell state)上进行空间平滑。

4.2.2 调节空间平滑任务权重

第2组实验是考察进一步调整权重之后，空间平滑方法的效果。对BLSTM网络最后一层的细胞状态$\boldsymbol{c}_t$进行空间平滑，调节权重，得到表2的结果。实验选择的权重分别为：0.01, 0.001, 0.0009, 0.0008, 0.0007。分析表2可以得出以下结论，当权重取得0.0009时，空间平滑方法的WER下降最明显，权重增大或减小时，性能都有所降低。最好的结果在CallHm子集上错误率下降0.7%，Swbd子集上降低0.5%，总集上绝对降低0.6%，相对下降4%。由此可以得出，本文提出的对BLSTM模型的一部分值进行空间平滑的方法可以提升网络的性能，降低语音识别的WER，一定程度上减轻了网络的过拟合问题。

4.2.3 添加L2正则

为了进一步验证文中提出的空间平滑方法的有效性，以及和其他常用正则方法的互补能力，本文

表 2 不同权重下的细胞状态值c_t的空间平滑结果

空间平滑权重(c)	CallHm WER (%)	Swbd WER (%)	总计WER (%)
无	20.0	10.3	15.2
0.0100	20.3	10.4	15.4
0.0010	19.7	9.8	14.8
0.0009	**19.3**	**9.8**	**14.6**
0.0008	19.6	9.7	14.7
0.0007	19.9	9.8	15.0

进一步分析了空间平滑方法与L2正则化方法的融合性能。实验结果见表3。这里的空间平滑是在细胞值上进行的，权值为0.0009，错误率下降0.6%。当只增加L2正则时，错误率在基线的基础上下降了0.9%，如果同时在网络中使用L2正则和空间平滑方法，识别结果的错误率能够下降绝对1.3%，相对降低8.6%。由此可见，本文提出的空间平滑方法与L2正则有着互补的作用，可以配合使用来缓解BLSTM网络的过拟合问题，使得模型达到更高的识别准确率。

表 3　网络中添加L2正则后的结果

L2正则 有/无	空间平滑 有/无	CallHm WER (%)	Swbd WER (%)	总计WER (%)
无	无	20.0	10.3	15.2
无	有	19.3	9.8	14.6
有	无	19.0	9.5	14.3
有	有	**18.5**	**9.3**	**13.9**

5　结束语

本文针对序列化准则下BLSTM模型容易过拟合的问题，提出了一种空间平滑方法，将空间平滑作为网络训练目标的一个正则项，同时探究了不同的权重、不同的平滑位置对识别性能的影响。在电话交谈语音识别任务上的实验结果表明，文中提出的空间平滑的方法使得识别WER相对下降了4%，结合L2正则之后，错误率相对降低了8.6%。由此可见，空间平滑是一种有效地减轻BLSTM网络过拟合的方法，它可以和L2正则互补地提升网络性能。下一步的研究会尝试使用一些其他的目标模式，让输出网格去和它匹配，来进行网络激活值的正则化，并且希望能够获得进一步的性能提升。

参 考 文 献

[1] LI X, and WU X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 2015: 4520–4524. doi: 10.1109/ICASSP.2015.7178826.

[2] CHEN K and HUO Q. Training deep bidirectional LSTM acoustic model for LVCSR by a context-sensitive-chunk BPTT approach[J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing* (*TASLP*), 2016, 24(7): 1185–1193. doi: 10.1109/TASLP.2016.2539499.

[3] AXELROD S, GOEL V, Gopinath R, *et al*. Discriminative estimation of subspace constrained gaussian mixture models for speech recognition[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, 15(1): 172–189. doi: 10.1109/TASL.2006.872617.

[4] POVEY D, KANEVSKY D, KINGSBURY B, *et al*. Boosted MMI for model and feature-space discriminative training[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA, 2008: 4057–4060. doi: 10.1109/ICASSP.2008.4518545.

[5] POVEY D and KINGSBURY B. Evaluation of proposed modifications to MPE for large scale discriminative training[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, USA, 2007: 321–324. doi: 10.1109/ICASSP.2007.366914.

[6] HUANG Z, SINISCALCHI S M, and LEE C H. Hierarchical Bayesian combination of plug-in maximum a posteriori decoders in deep neural networks-based speech recognition and speaker adaptation[J]. *Pattern Recognition Letters*, 2017, 98(15): 1–7. doi: 10.1016/j.patrec.2017.08.001.

[7] POVEY D. Discriminative training for large vocabulary speech recognition[D].[Ph.D. dissertation], University of Cambridge, 2003.

[8] ZHOU P, JIANG H, DAI L R, *et al*. State-clustering based multiple deep neural networks modeling approach for speech recognition[J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing* (*TASLP*), 2015, 23(4): 631–642. doi: 10.1109/TASLP.2015.2392944.

[9] SRIVASTAVA N, HINTON G, KRIZHEYSKY A, *et al*. Dropout: A simple way to prevent neural networks from overfitting[J]. *The Journal of Machine Learning Research*, 2014, 15(1): 1929–1958.

[10] GOODFELLOW I, BENGIO Y, and COURVILLE A, Deep Learning[M], Cambridge, MA: MIT Press, 2016: 228–230.

[11] POVEY D, PEDDINTI V, GALVEZ D, *et al*. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]. International Speech Communication Association (INTERSPEECH), San Francisco, USA, 2016: 2751–2755. doi: 10.21437/Interspeech.2016-595.

[12] SAHRAEIAN R, and VAN D. Cross-entropy training of DNN ensemble acoustic models for low-resource ASR[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, 26(11): 1991–2001. doi: 10.1109/TASLP.2018.2851145.

[13] LIU P, LIU C, JIANG H, *et al*. A constrained line search optimization method for discriminative training of HMMs[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2008, 16(5): 900–909. doi: 10.1109/TASL.2008.925882.

[14] WU C, KARANASOU P, GALES M J, *et al*. Stimulated

deep neural network for speech recognition[C]. International Speech Communication Association (INTERSPEECH), San Francisco, USA, 2016: 400–404. doi: 10.21437/Interspeech.2016-580.

[15] Wu C, CALES M J F, RAGNI A, *et al*. Improving interpretability and regularization in deep learning[J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing* (*TASLP*), 2018, 26(2): 256–265. doi: 10.1109/TASLP.2017.2774919.

[16] KO T, PEDDINTI V, POVEY D, *et al*. Audio augmentation for speech recognition[C]. International Speech Communication Association (INTERSPEECH), Dresden, Germany, 2015: 3586–3589. doi: 10.21437/Interspeech.2015-571.

[17] LAURENT C, PEREYRA G, BRAKEL P, *et al*. Batch normalized recurrent neural networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016: 2657–2661. doi: 10.1109/ICASSP.2016.7472159.

李文洁：女，1993年生，博士生，研究方向为语音信号处理、语音识别、声学模型、远场语音识别等.

葛凤培：女，1982年生，副研究员，研究方向为语音识别、发音质量评估、声学建模及自适应等.

张鹏远：男，1978年生，研究员，硕士生导师，研究方向为大词表非特定人连续语音识别、关键词检索、声学模型、鲁棒语音识别等.

颜永红：男，1967年生，研究员，博士生导师，研究方向为语音信号处理、语音识别、口语系统及多模系统、人机界面技术等.