应用声学
第 37 卷 第 6 期
2018 年 11 月
Journal of Applied Acoustics
Vol. 37, No.6
November, 2018

◇ 研究报告 ◇

基于改进卷积神经网络算法的语音识别

杨 洋† 汪毓铎

(北京信息科技大学信息与通信工程学院 北京 100101)

摘要 为了解决传统卷积神经网络识别连续语音数据时识别性能较差的问题，提出一种改进的卷积神经网络算法。该方法引入Fisher准则以及L2正则化约束，在反向传播调整参数阶段，既保证参数误差的最小化，又确保分类以后的样本类间分布较分散，类内分布较集中，同时保证网络权值具有合适的数量级以有效缓解过拟合问题；采用一种更符合生物神经元激活特性的新型log激活函数进行卷积神经网络的优化，进一步提高语音识别的正确率。在语音识别库TIMIT以及THCHS30上的实验结果表明，相较于传统卷积神经网络算法，该文提出的改进算法能较好地提高语音识别率，且泛化能力更强。

关键词 语音识别，卷积神经网络，Fisher准则，L2正则化，log激活函数

中图法分类号： TN912.3 **文献标识码：** A **文章编号：** 1000-310X(2018)06-0940-07

DOI: 10.11684/j.issn.1000-310X.2018.06.016

Speech recognition based on improved convolutional neural network algorithm

YANG Yang WANG Yuduo

(*School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing* 100101, *China*)

Abstract An improved convolutional neural network (CNN) algorithm is proposed to solve the problem of poor recognition performance when the traditional CNN identifies continuous speech corpus. In this method, Fisher criterion and L2 regularization constraint are introduced. In the phase of back propagation adjustment parameters, it not only ensures the minimum of parameter errors, but also ensures that the distribution of samples after classification is more scattered, and the distribution within class is more concentrated. At the same time, the weights of the network are guaranteed to have the appropriate order of magnitude to effectively alleviate the problem of over-fitting. In order to further improve the accuracy of speech recognition, a new log activation function which is more consistent with the biological neuron is used to optimize the CNN. Experiments on speech corpus TIMIT and THCHS30 show that compared with the traditional CNN algorithm, the improved algorithm proposed in this paper can better improve the accuracy and the generalization ability.

Key words Speech recognition, Convolutional neural network, Fisher criterion, L2 regularization, log activation function

2018-01-25 收稿; 2018-05-01 定稿

作者简介: 杨洋 (1994-), 女, 河南商丘人, 硕士研究生, 研究方向: 语音信号处理。

† 通讯作者 E-mail: 18811536735@163.com

1 引言

自动语音识别 (Automatic speech recognition, ASR) 技术能够使人与人、人与机器实现更顺畅的交流 [1]。语音识别技术经过 50 多年的发展，为人们的生活带来了巨大的变化，比如语音智能控制家居设备以及车载娱乐等。语音识别中两种典型的并且截止到现在仍被广泛使用的方法有 (1) 基于高斯混合模型-隐马尔可夫模型 (Gaussian mixture model-Hidden Markov model, GMM-HMM) 的语音识别系统；(2) 基于深度学习-隐马尔可夫模型 (Deep learning-Hidden Markov model, DL-HMM) 的语音识别系统 [2]。传统的 GMM-HMM 方法在扁平浅层生成式模型的基础上，结合线性判别分析 (Linear discriminant analysis, LDA)、最大似然训练准则 (Maximum likelihood estimation, MLE)[3] 以及说话人自适应等技术，在简单的场景中得到了较好的运用。但是随着技术的发展以及人类需求的不断提高，需要应用自动语音识别的场景越来越复杂，传统的 GMM-HMM 已不再适用，具有更加强大建模能力的声学模型成为迫切的需要，由此基于 DL-HMM 声学模型的语音识别系统开始流行。

当前识别语音信号主流的深度学习算法为深层神经网络 (Deep neural network, DNN)、长短时记忆网络 (Long short-term memory, LSTM) 以及卷积神经网络 (Convolutional neural network, CNN)。2009 年，DNN 首次被用于加强隐马尔可夫声学模型的构建，对 TIMIT 语音数据库进行音素级的识别 [4]，识别效果得到很大改善。文献 [5] 利用 CNN、LSTM 和 DNN 的互补性将它们组合成一个统一的 CLDNN 体系结构，使得语音识别率得到 4%~6% 的相对改善。目前，IBM、微软、百度等多家机构相继推出了自己的深度学习语音识别模型，使得语音识别研究取得了很大的突破。其中 CNN 是一种深度结构学习算法，相较于其他深层神经网络结构，CNN 具有权值共享、局部卷积以及池化的明显特征，这些特点决定了 CNN 具有复杂度低的特性 [6]。因此，在语音识别领域，研究者开始将目光转向 CNN，构建 CNN-HMM 声学模型。与只使用 DNN-HMM 的声学模型相比，文献 [7] 提出的 CNN-DNN-HMM 结构在大词汇量连续语音识别中获得了更高的识别正确率。文献 [8] 针对特定情形下的语音识别，比如距离较远的语音识别系统，提出的 CNN 语音识别模型要比 DNN 更有效，适应能力更强。很长一段时间以来 CNN 都是与其他深层神经网络结合，也就是在底层采用 CNN，高层采用 DNN 等其他的深度神经网络模型。然而在最近的一些研究工作中，CNN 不再只应用在底层，文献 [9–12] 在进行语音识别时，采用大于 10 层的非常深的 CNN 模型，极大地提高了系统性能。

CNN 在卷积层采用局部连接、共享权值的方式提取特征，减少了权值的训练数量并能在一定程度上防止过拟合问题的出现，在卷积层之后经过池化层 (又叫聚合层) 的最大池化或平均池化技术的处理，使得模型结构进一步简化，并能增强语音识别系统的鲁棒性。本文基于标准语音数据库 TIMIT 以及 THCHS30 提出一种改进的 CNN 算法，在反向传播调节参数阶段，采用结合 Fisher 准则以及 L2 正则化的约束项，既保证参数误差的最小化，又同时使得不同类型的样本在分类以后相对分散，类内样本间相对集中，从而使得训练的参数更接近于最优值以及减轻语音识别容易出现的过拟合问题，并采用一种更符合生物神经元的新型的 log 激活函数进行 CNN 的优化，进一步降低语音识别的错误率。

2 深度卷积模型基本原理

CNN 的结构如图 1 所示。深度卷积模型一般包括输入层、卷积层、池化层、全连接层以及输出层，卷积层和池化层是特殊的隐含层 [13]。通常卷积层之后是聚合层，两者以一组或多组的形式成对出现，但是特定场景下也可以采用隐含层不包含聚合层的特殊深度卷积模型。全连接层可以为单层也可以为多层，其作用就是将经过池化处理以后的信号特征进行全连接，然后送到输出层进行分类，输出层的激活单元一般选择 softmax 函数。一个卷积层包括多个卷积特征图谱 (又叫卷积面)，每个卷积特征图谱对应于一个卷积滤波器 (又叫卷积核)，通过对应卷积核对输入的信号特征进行局部的过滤可得到该卷积特征图谱上的神经元输入。池化层以固定的窗口大小对每个卷积特征图谱做下采样，一般都采用最大池化技术，也就是取每个卷积面在池化窗口大小内的最大值作为对应下采样面的神经元输入，这种方式明显降低了每个下采样面的神经元数目。

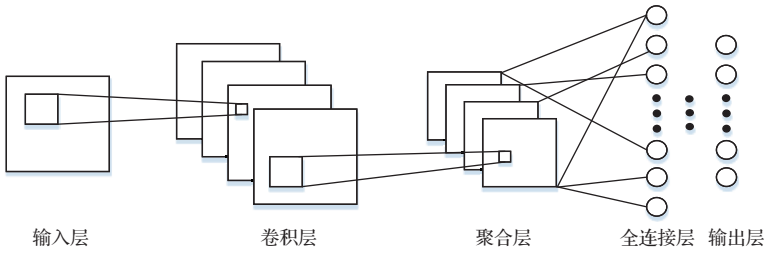

图 1 标准卷积模型

Fig. 1 Standard convolution model

近几年，图像识别、目标定位等领域被研究者广泛使用的深度学习算法就是CNN。CNN用于语音识别时，提取的声学特征仍然采用类似于图像识别中的二维矩阵输入形式，一个维度代表时域，一个维度代表频域[14]。声学特征的二维矩阵输入形式如图2所示，假设语音数据被分成25帧，25帧语音数据的静态声学特征、一阶差分和二阶差分沿水平方向的时间域(语音数据帧)和垂直方向的频率(频带指数)分布。将声学特征二维映射矩阵输入到CNN，可进行二维卷积运算提取深层次的语音信号特征。

标准CNN的代价函数一般都是最小均方误差函数，CNN在训练和学习参数的过程中要使得该代价函数达到最小。在前向传播特征学习阶段，CNN依据局部卷积、权值共享以及下采样原则，使得模型结构复杂度大大降低，鲁棒性增强。在微调阶段，通过误差反向传导算法(Back propagation，BP)[15]自顶向下小幅度地调节所有层的权值和偏置，使得输出层每个单元的真实输出值与输入样本标签值最接近。

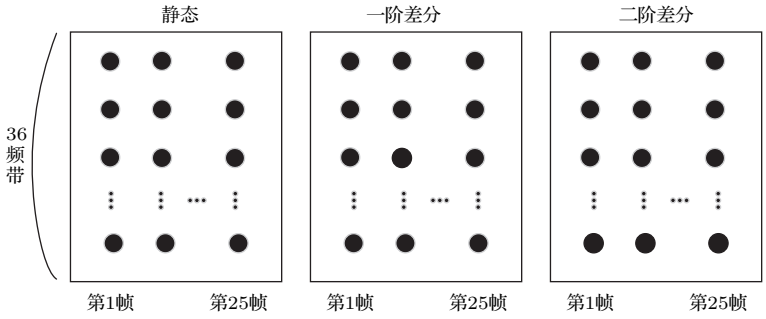

图 2 声学特征二维映射矩阵

Fig. 2 Acoustic characteristics two-dimensional mapping matrix

3 深度卷积模型的优化

识别连续语音库时代价函数只考虑最小均方误差函数是比较单一的。一般会将总数据集按照一定的规则进行划分，一部分语音数据用来训练CNN，一部分数据用来测试已经训练好的CNN模型的性能。训练数据量太小，模型将过度学习数据，容易过度拟合，因此引入L2正则化的约束；而且为了实现最优分类，使得最终的样本分类结果可以达到类内距离小、类间距离大的目的，本文同时也会引入Fisher准则的约束，与L2正则化结合使用，使得训练的网络权值和偏置与最优值更加接近。

3.1 深度卷积模型基于改进代价函数的BP算法

设训练集$\mathbf{S}=\left\{\left(x^i,y^i\right),1\leqslant i\leqslant m\right\}$，训练集$\mathbf{S}$中包含$m$个样本$\left\{x^1,x^2,\cdots,x^m\right\}$，它们可被划分为$n$个类别，$y^i$是样本$x^i$对应的类别标签值。则CNN的最小平方误差损失函数为

$$E=\frac{1}{m}\left[\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}(x^{(i)})-y^{(i)}\right\|_2^2\right)\right], \quad (1)$$

式(1)中，E为最小平方误差损失函数，m表示总的样本数，$h_{W,b}\left(x^{(i)}\right)$为训练样本$x^i$经过CNN训练后得到的实际输出，$W$表示各层神经元之间连接的权

值，b 表示对应的偏置。

Fisher 准则在提取语音信号的整体特征时是以训练语音数据集的音素类别信息为基础的。其基本原理是基于类内离散程度矩阵和类间离散程度矩阵，依据一定的数学计算规则寻找一个最佳投影空间，在该投影空间上样本点尽量按类别区分开，从而实现最佳分类并缩小了特征空间的维数[16]。小规模数据很容易出现过拟合问题，L2 正则化在代价函数的基础上加上一个正则化惩罚项，减小权值的数量级，限制过拟合。

借鉴 Fisher 准则和 L2 正则化的思想，加入类内和类间散布度量函数和 L2 正则化惩罚项的代价函数表示为

$$\begin{aligned} C &= E + aJ_w - bJ_b + \frac{\lambda}{2}\sum_w \|w\|_2^2 \\ &= C_0 + \frac{\lambda}{2}\sum_w \|w\|_2^2, \end{aligned} \tag{2}$$

$$J_w = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{m_i}\left\|h_{W,b}\left(x^{(i,j)}\right) - M^{(j)}\right\|^2, \tag{3}$$

$$J_b = \frac{1}{2}\sum_{k=1}^{n}\sum_{j=k+1}^{n}\left\|M^{(k)} - M^{(j)}\right\|^2, \tag{4}$$

其中，J_w 为类内散布度量函数；J_b 为类间散布度量函数；a、b 为常数，取值范围通常在 $0 \sim 1$；$\lambda(\lambda > 0)$ 为正则化参数，用来权衡正则项与 C_0 的比重；w 为各层神经元连接权值。

J_w 定义为训练样本的真实输出与其所属类的样本均数之间的距离总和，其中 m_i 为第 j 类的样本数量，样本的种类数为 n，第 j 类的第 i 个样本的实际输出为 $h_{W,b}\left(x^{(i,j)}\right)$；$J_b$ 定义为所有异类样本均数的距离总和，$M^{(k)}$ 和 $M^{(j)}$ 分别为第 k 类和第 j 类的样本均值，第 j 类的样本均值 $M^{(j)}$ 为

$$M^{(j)} = \frac{\sum_{i=1}^{m_i} h_{W,b}\left(x^{(i,j)}\right)}{m_i}. \tag{5}$$

CNN 在用 BP 算法进行参数微调时，最重要的就是利用代价函数计算出输出层的反传误差信号(残差)，然后将残差由输出层自顶向下的传播至输入层，利用梯度下降算法进行权值和偏置的更新。

对于最小平方误差函数，输出层每个单元的反向传播残差计算公式为

$$\begin{aligned} \frac{\partial E}{\partial u_R^i} &= \frac{\partial\left[\frac{1}{2}\left\|h_{W,b}(x^{(i)}) - y^{(i)}\right\|_2^2\right]}{\partial u_R^i} \\ &= \left(f\left(u_R^i\right) - y^{(i)}\right) \circ f'\left(u_R^i\right), \end{aligned} \tag{6}$$

$$u_R^i = w_R^i x_{R-1}^i + b_R^i, \tag{7}$$

其中，u_R^i 为输出层第 i 个单元的输入，w_R^i 和 b_R^i 分别为输出层 R 第 i 个单元的权重和偏置，x_{R-1}^i 表示全接入层 $R-1$ 层第 i 个单元的输出，$f(\bullet)$ 为激活函数。

对于 J_w 类内散布度量函数，输出层的反传误差信号为

$$\begin{aligned} \frac{\partial J_w}{\partial u_R^i} &= \frac{\partial}{\partial u_R^i}\left(\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{m_i}\left\|h_{W,b}(x^{(i,j)}) - M^{(j)}\right\|^2\right) \\ &= \sum_{j=1}^{n}\sum_{m=1}^{m_i}\left(h_{W,b}(x^{(i,j)}) - M^{(j)}\right) \\ &\quad \times\left((h_{W,b}(x^{(i,j)}))' - (M^{(j)})'\right) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{m_i}\left(f\left(u_R^i\right) - \frac{1}{m_i}f\left(u_R^i\right)\right) \\ &\quad \times\left(f'\left(u_R^i\right) - \frac{1}{m_i}f'\left(u_R^i\right)\right). \end{aligned} \tag{8}$$

对于 J_b 类间散布度量函数，输出层的反向传播残差为

$$\begin{aligned} \frac{\partial J_b}{\partial u_R^i} &= \frac{\partial}{\partial u_R^i}\left(\frac{1}{2}\sum_{k=1}^{n}\sum_{j=i+1}^{n}\left\|M^{(k)} - M^{(j)}\right\|^2\right) \\ &= \sum_{k=1}^{n}\sum_{j=i+1}^{n}\left(M^{(k)} - M^{(j)}\right) \\ &\quad \times\left((M^{(k)})' - (M^{(j)})'\right) \\ &= \sum_{k=1}^{n}\sum_{j=i+1}^{n}\left(M^{(k)} - M^{(j)}\right) \\ &\quad \times\left[m_i M^{(k)} \cdot \left(\frac{1}{m_i} - M^{(k)}\right)\right. \\ &\quad \left. - m_i M^{(j)} \cdot \left(\frac{1}{m_i} - M^{(j)}\right)\right]. \end{aligned} \tag{9}$$

则输出层第 i 个单元的反传残差为

$$\delta_R^i = \frac{\partial C_0}{\partial u_R^i} = \frac{\partial E}{\partial u_R^i} + a\frac{\partial J_w}{\partial u_R^i} - b\frac{\partial J_b}{\partial u_R^i}. \tag{10}$$

计算出输出层的反向传播残差以后，通过 BP 算法每次迭代更新网络参数时，能使参数向更有利于分类的方向靠拢。加上正则项以后，梯度下降法更新所有网络参数的计算公式为

$$\begin{cases} w \to w - \eta\dfrac{\partial C}{\partial w} \\ \qquad = w - \eta\dfrac{\partial C_0}{\partial w} - \eta\lambda w \\ \qquad = (1-\eta\lambda)w - \eta\dfrac{\partial C_0}{\partial w}, \\ b \to b - \eta\dfrac{\partial C}{\partial b} = b - \eta\dfrac{\partial C_0}{\partial b}. \end{cases} \tag{11}$$

式 (2) 中的参数需由实验确定，本文选择 $a = 0.03$，$b = 0.03$，$\lambda = 0.0004$。

3.2 改进的 log 激活函数

由于线性激活函数的复杂性有限，从数据中学习复杂特征的能力较弱，因此在 CNN 中一般均采用非线性激活函数，如 sigmoid 函数、tanh 函数，表达式分别如式 (12) 和式 (13) 所示：

$$\text{sigmoid}\,(x) = \frac{1}{1+\mathrm{e}^{-x}}, \tag{12}$$

$$\tanh(x) = \frac{1-\mathrm{e}^{-2x}}{1+\mathrm{e}^{-2x}}. \tag{13}$$

sigmoid 函数和 tanh 函数均是饱和的，梯度在向底层传递时很容易消失，而且 tanh 函数关于原点对称，这与生物神经元的激活特征是不相符的。根据大脑神经元激活的仿真模型，提出了一种新的 log 激活函数来优化 CNN，使得语音识别的词错率得到进一步的降低，表达式如式 (14) 所示：

$$f(x) = \begin{cases} \ln(x+1), & x > 0, \\ 0, & x \leqslant 0. \end{cases} \tag{14}$$

新型 log 激活模型的函数图像如图 3 所示。

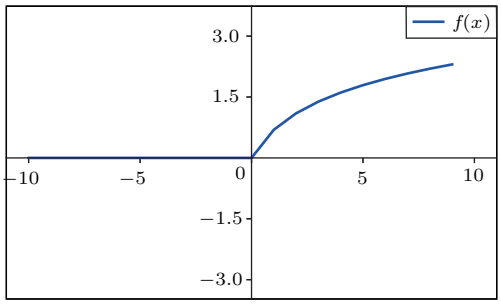

图 3 新型 log 激活函数

Fig. 3 New log activation function

当神经元输入特征值小于零时，新型 log 激活函数将输出值强制为零，符合生物学神经元的稀疏激活特性，缓解过拟合问题的发生；输入特征值大于零时，输出值随输入值呈非线性递增变化，能有效缓解梯度消失的问题。

4 实验结果与分析

为了验证所提算法的有效性，本文分别基于 TIMIT 和 THCHS30 数据库进行实验。由于连续语音识别最终识别出来的一般都是按照特定顺序排列的一串词，因此实验中采用的评测标准是词错率 (Word error rate, WER)，WER 值的大小与系统的整体性能优劣成反比。选用 n-gram 的统计语言模型，即当前词出现的概率只与其前面的 n 个词有关。

4.1 基于 TIMIT 数据库的实验

TIMIT 是常用的英文语音库，将数据库中的 462 个说话人的语音作为训练集，将 40 个说话人的语音作为测试集，两个集之间无重叠。选用 2-gram 的统计语言模型。在特征提取部分，广泛使用的声学特征是梅尔频率倒谱系数 (Mel-frequency cepstral coefficients, MFCC)，但是在提取 MFCC 对梅尔能量做离散余弦变换时会使能量值发生偏置，不利于 CNN 对特征做局部提取，因此实验选用帧与帧之间具有较强关联性的 FBANK 声学特征，每帧语音数据提取的 FBANK 特征维度为 36 维，同时对其做一阶和二阶差分扩展。对每帧声学特征做倒谱均值和归一化。实验中，沿时间轴左右各展开 5 帧，构成上下文相关的 11 帧串联长时特征。

本实验所训练的二维 CNN 的隐含层包括一层卷积层、一层聚合层和一层全连接层，输出采用 softmax 层。卷积核的数目为 256，卷积核的大小为 9×9，步进为 1×1；聚合层采用最大聚合算法，4 个神经元中选择最大的一个节点值作为输出，步长为 1×4；全连接层的神经元数为 1024。CNN 的输出层输出的是语音帧属于某个类别的后验概率，在语音识别中类指的是音素，实验中语音数据的音素类别数为 144(48 个音素，每个音素三个状态)。

表 1 给出了 TIMIT 测试集上不同模型之间的对比实验结果：Fisher 模型较一般模型，词错率降低了 1.1%，语音识别性能有较好的提升；在使用 L2 正则化进行改进以后，词错率有略微下降；而在进一步使用 log 激活函数进行优化以后，词错率又有略微下降。总体来说，使用优化 CNN 算法的识别正确率比传统 CNN 提高了 1.6%。由此可知，本文所提出的 CNN 改进算法在 TIMIT 语音数据库中能较好地提升语音识别的准确率。

表 1　TIMIT 测试集上不同模型之间的性能对比

Table 1　Performance comparison between different models on TIMIT corpus

模型	激活函数类型	代价函数	WER(%)
一般模型	sigmoid	平方误差	24.1
Fisher 模型	sigmoid	平方误差 + Fisher	23.0
Fisher 模型 + L2 正则化	sigmoid	平方误差 + Fisher + L2	22.7
Fisher 模型 + L2 正则化 + log 激活函数	log 激活函数	平方误差 + Fisher +L2	22.5

4.2　基于 THCHS30 数据库的实验

上述的 TIMIT 语音识别数据库是一个开源的英文语音库，为了验证所提算法的有效性，进一步在中文语音库上进行对比实验。实验在清华大学中文语音数据集 THCHS30 上进行，该语音库是总时长超过 30 h 的汉语普通话数据集，采样频率 16 kHz，采样大小 16 bits。其中训练数据集时长约为 25 h，共 10000 句；测试数据共 2495 句，全集长度约 6 h。选用 3-gram 的 word 级统计语言模型。

本实验采用四隐层的 CNN 结构，特征是 40 维的 FBANK，并且相邻的帧由 11 帧窗口 (每侧 5 个窗口) 连接。对每帧声学特征做倒谱均值和归一化以获得 CNN 的输入。其中，CNN 的第一个卷积层的卷积核尺寸为 9×9，卷积核数目为 128，步进为 1×1，聚合层采用最大聚合算法，步长为 3×3；第二个卷积层的卷积核尺寸为 4×4，卷积核数目为 256，步进为 1×1，聚合层步长为 2×2；第三层和第四层为全连接层，每个层由 1024 个单元组成；softmax 输出层由 3386 个单元组成。

表 2 给出了 THCHS30 测试集上不同模型之间的对比实验结果：WER 的变化趋势与 TIMIT 数据库的实验结果类似。总体来说，使用本文提出的优化 CNN 算法的语音识别率比一般模型提高了 1.49%。从表 1 和表 2 的性能对比实验结果中可看出，本文所提出的改进 CNN 算法无论是基于英文语音库还是中文语音库，识别性能均比传统的 CNN 模型要好。

表 2　THCHS30 测试集上不同模型之间的性能对比

Table 2　Performance comparison between different models on THCHS30 corpus

模型	激活函数类型	代价函数	WER(%)
一般模型	sigmoid	平方误差	23.68
Fisher 模型	sigmoid	平方误差 + Fisher	22.76
Fisher 模型 + L2 正则化	sigmoid	平方误差 + Fisher + L2	22.43
Fisher 模型 + L2 正则化 + log 激活函数	log 激活函数	平方误差 + Fisher +L2	22.19

5　结论

本文提出的改进 CNN 算法将 Fisher 准则和 L2 正则化作为惩罚项引入 CNN 的代价函数中，并基于生物神经元的激活模型提出一种新型的 log 激活函数，有效地改善了 CNN 的语音识别性能。在 TIMIT 以及 THCHS30 语音库上的实验结果表明，所提算法较好地缓解了语音识别时容易出现的过拟合问题，并使得各类音素状态间的距离大，音素状态内的语音数据帧距离小，相比于标准 CNN 网络，改进的 CNN 算法的泛化能力更强，语音识别率更高。

参　考　文　献

[1] 俞栋, 邓力, 俞凯, 等. 解析深度学习语音识别实践 [M]. 北京: 电子工业出版社, 2016.

[2] 侯一民, 周慧琼, 王政一. 深度学习在语音识别中的研究进展综述 [J]. 计算机应用研究, 2017, 34(8): 2241–2246.

Hou Yimin, Zhou Huiqiong, Wang Zhengyi. Overview of speech recognition based on deep leaarning[J]. Application Research of Computers, 2017, 34(8): 2241–2246.

[3] Li J, Mohamed A, Zweig G, et al. LSTM time and frequency recurrence for automatic speech recognition[C]//Automatic Speech Recognition and Understanding, IEEE, 2016: 187–191.

[4] Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]//Nips Workshop on Deep Learning for Speech Recognition and Related Application, 2009: 39.

[5] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2015: 4580–4584.

[6] 孙艳丰, 齐光磊, 胡永利, 等. 基于改进 Fisher 准则的深度卷积神经网络识别算法 [J]. 北京工业大学学报, 2015, 41(6): 835–841.
Sun Yanfeng, Qi Guanglei, Hu Yongli, et al. Deep convolution neural network recognition algorithm based on improved Fisher criterion[J]. Journal of Beijing University of Technology, 2015, 41(6): 835–841.

[7] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013: 8614–8618.

[8] Huang J T, Li J, Gong Y. An analysis of convolutional neural networks for speech recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2015: 4989–4993.

[9] Sercu T, Puhrsch C, Kingsbury B, et al. Very deep multilingual convolutional neural networks for LVCSR[C]//Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 4955–4959.

[10] Sainath T N, Kingsbury B, Saon G, et al. Deep convolutional neural networks for large-scale speech tasks[J]. Neural Networks, 2015, 64: 39–48.

[11] Yu D, Xiong W, Droppo J, et al. Deep convolutional neural networks with layer-wise context expansion and attention[C]// Interspeech, 2016: 17–21.

[12] Qian Y, Bi M, Tan T, et al. Very deep convolutional neural networks for noise robust speech recognition[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(12): 2263–2276.

[13] 梁玉龙, 屈丹, 李真, 等. 基于卷积神经网络的维吾尔语语音识别 [J]. 信息工程大学学报, 2017, 18(1): 44–50.
Liang Yulong, Qu Dan, Li Zhen, et al. Uyghur speech recognition based on convolutional neural network[J]. Journal of Information Engineering University, 2017, 18(1): 44–50.

[14] 黄玉蕾, 罗晓霞, 刘笃仁. MFSC 系数特征局部有限权重共享 CNN 语音识别 [J]. 控制工程, 2017, 24(7): 1507–1513.
Huang Yulei, Luo Xiaoxia, Liu Duren. Local finite weight sharing of MFSC coefficients based CNN speech recognition[J]. Control Engineering of China, 2017, 24(7): 1507–1513.

[15] Hori T, Watanabe S, Zhang Y, et al. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM[J]. arXiv preprint arXiv: 1706.02737, 2017.

[16] Zeiler S, Nicheli R, Ma N, et al. Robust audiovisual speech recognition using noise-adaptive linear discriminant analysis[C]//IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2016: 2797–2801.