ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 331－337
DOI:10.16337/j.1004-9037.2016.02.013

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于区分性准则的 Bottleneck 特征及其在 LVCSR 中的应用

刘迪源　郭　武

(中国科学技术大学语音及语言信息处理国家工程实验室,合肥,230027)

摘　要:基于深层神经网络中间层的 Bottleneck(BN)特征由于可以采用传统的混合高斯模型-隐马尔可夫建模(Gaussian mixture model-hidden Markov model, GMM-HMM),在大规模连续语音识别中获得了广泛的应用。为了提取区分性的 BN 特征,本文提出在使用传统的 BN 特征训练好 GMM-HMM 模型之后,利用最小音素错误率(Minimum phone error, MPE)准则来优化 BN 网络参数以及 GMM-HMM 模型参数。该算法相对于其他区分性训练算法而言,采用的是全部数据作为一个大的数据包,而不是小的包方式来训练深度神经网络,从而可以大大加快训练速度。实验结果表明,优化后的 BN 特征提取网络比传统方法能获得 9%的相对词错误率下降。

关键词:语音识别;神经网络;区分性训练;Bottleneck 特征

中图分类号:TN912.34　　**文献标志码**:A

Discriminative Criterion Based Bottleneck Feature and Its Application in LVCSR

Liu Diyuan, Guo Wu

(National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China)

Abstract: Bottleneck (BN) features based on the middle layer of deep neural network have been widly applicated to large vocabulary continuous speech recognition (LVCSR), because they can use the traditional Gaussian mixture density hidden Markov model (GMM-HMM) for acoustic modeling. In order to extract discriminative bottleneck features, the parameters of the BN feature extractor and GMM-HMM are optimized jointly by using the minimum phone error (MPE) criterion after training the GMM-HMM using the conventional BN features. Different from other discriminative training method, large batches instead of mini-batch in conventional neural network optimization are used to obtain the statistics, which accelerates training speed. Experiments demonstrate that the proposed bottleneck feature extractor can outperform the traditional methods with 9% relative word error reduction.

Key words: speech recognition; neural networks; discriminative training; Bottleneck feature

引　　言

在基于混合高斯模型和隐马尔科夫模型的自动语音识别中,区分性训练技术可以提高混合高斯模型-隐马尔可夫建模(Gaussian mixture model-hidden Markov model, GMM-HMM)的语音识别率。区分

收稿日期:2014-03-03;**修订日期**:2014-04-14

性训练可以应用在模型空间中，例如最大互信息准则(Maximum mutual information, MMI)[1]、最小分类错误率准则(Minimum classification error, MCE)[2,3]以及最小音素错误率准则(Minimum phone error, MPE)[4]。同样，区分性训练也可以应用在特征空间中，比如特征域最小词错误率准则(feature Minimum phone error, fMPE)[5]，特征域最大互信息量准则(feature boosted Maximum mutual information, fbMMI)[6]，神经网络-特征域最大互信息准则(Neural network-feature minimum phone error, NN-fMMI)[7]。

神经网络在人机语音交互中有广泛的应用[8,9]。近几年，随着深层神经网络(Deep neural network, DNN)在语音识别中的成功应用[10]，语音识别的性能得到了很大提高。DNN 网络的参数(权值和偏差)通常分两步来训练：首先使用受限玻尔兹曼机(Restricted Boltzmann machine, RBM)的预训练来逐层初始化 DNN 网络参数，然后利用基于交叉熵准则的误差反向传播(Back propagation, BP)算法训练网络。与 GMM-HMM 中的区分性训练类似，神经网络序列-区分性训练[11,12]技术应用到了 DNN 训练中。序列-区分性训练通过使用 MMI，MPE 等区分性准则代替传统的交叉熵准则来训练 DNN 网络，进一步提高了语音识别率。

神经网络的另一个重要应用是 Bottleneck 特征提取[13-15]。Bottleneck 特征一般通过如下方法来提取：首先使用传统方法训练一个多层神经网络，该网络具有一个节点数很少的中间隐层(BN 层)；然后去掉 BN 层之后的网络，使用剩下的 BN 特征提取网络来提取 BN 特征；最后使用得到的 BN 特征训练传统的 GMM-HMM 模型。基于 BN 特征的 GMM-HMM 有时甚至能获得比 DNN 更好的识别性能，BN 特征提取网络比 DNN 结构更小，能加快神经网络训练和语音解码的速度。

文献[16]提出了一种基于格的区分性 BN 特征提取方法。首先使用传统的方法训练好 BN 特征提取网络，在该网络之后增加一个 GMM 层；然后使用 MMI 准则来优化网络参数，在优化训练的过程中，GMM 模型的参数保持不变；最后使用优化后的 BN 特征提取网络来提取新的 BN 特征，重新训练 GMM 模型。本文使用 Mini-batch 的方式训练网络，难以实现分布式计算，同时，Mini-batch 的方式更新难以计算间接梯度部分，不会更新 GMM 模型参数。针对文献[16]的问题，本文提出另外一种提取区分性 BN 特征的方法：首先使用传统的交叉熵准则训练 BN 网络；然后使用 BN 特征提取网络提取 BN 特征，训练 GMM-HMM；最后使用 MPE 准则优化网络参数。与文献[16]不同的是，本文使用所有训练集数据作为一个大数据包来计算 MPE 目标函数对网络参数的梯度，可实现分布式计算。同时，由于大数据包模式的更新方式可以方便地计算间接梯度部分，在训练过程中 BN 特征提取网络参数和 GMM 参数将会同时更新，这使得模型参数估计得更加精确。实验结果表明，本文提出的两种方法都能很大提高语音识别率。

1 基线 Bottleneck 神经网络

图 1 是基线 Bottleneck 神经网络。此网络的结构和文献[13,16]描述的相似，它包括 7 层：输入层、输出层以及 5 个隐层。输入层有 429 个节点，通过声学特征扩展得到；隐层节点数分别为 2 048，2 048，39，2 048，2 048，中间隐层为 BN 层；输出层节点数和 Tri-phone 绑定状态数相同。网络参数通过 RBM pre-training 初始化，然后根据交叉熵准则，使用 1 024 帧的 Mini-batches 来训练网络。训练好 Bottleneck 神经网络移除 BN 层之后的网络，得到 BN 特征提取，如图 2 所示。图 2 中各个参数的关系为

$$\begin{aligned} \boldsymbol{v}_1 &= \boldsymbol{W}_1 \boldsymbol{x} \qquad \boldsymbol{h}_1 = \sigma(\boldsymbol{v}_1) \\ \boldsymbol{v}_2 &= \boldsymbol{W}_2 \boldsymbol{h}_1 \qquad \boldsymbol{h}_2 = \sigma(\boldsymbol{v}_2) \\ \boldsymbol{y} &= \boldsymbol{h}_3 = \boldsymbol{v}_3 = \boldsymbol{W}_3 \boldsymbol{h}_2 \end{aligned} \tag{1}$$

式中：$\boldsymbol{x}$ 为标准的声学特征；$\boldsymbol{y}$ 为线性输出的 BN 特征；$\boldsymbol{W}=\{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3\}$ 为 BN 特征提取网络的参数(权

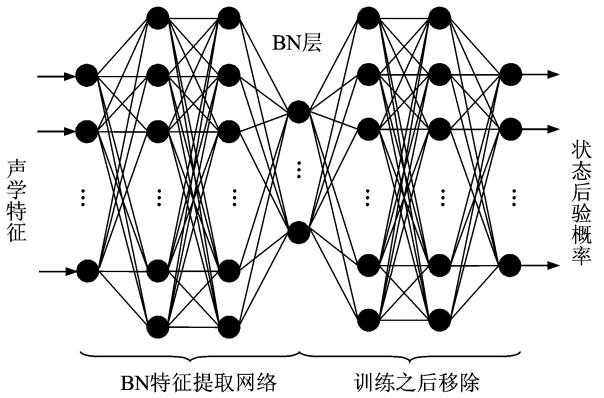

图 1 基线 Bottleneck 神经网络

Fig. 1 Baseline Bottleneck neural networks

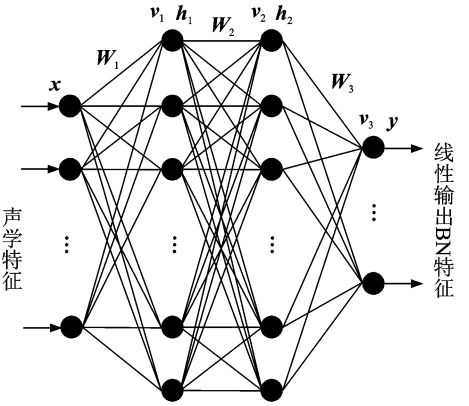

图 2 BN 特征提取网络

Fig. 2 BN feature extraction network

值和偏差);$\boldsymbol{W}_i$ 是大小为 $n_i(n_{i-1}+1)$的矩阵;n_i 为第 i 层的节点数;σ 为 Sigmoid 激活函数

$$\sigma(v)=\frac{1}{1+\mathrm{e}^{-v}} \tag{2}$$

2 区分性 BN 特征提取网络

2.1 训练流程

区分性 BN 特征提取网络的训练流程如下:

(1)训练基于感觉加权线性预测(Perceptual linear prediction, PLP)特征的 GMM-HMM 来获得训练集数据的帧级标签。依次使用最大似然准则和 MPE 准则训练 GMM-HMM 模型,再用该 GMM-HMM 模型对训练数据进行数据对齐,获得训练集数据的帧级标签。

(2)使用交叉熵准则训练 Bottleneck 神经网络。

(3)使用步骤(2)训练的神经网络提取 BN 特征,用这些 BN 特征来训练 GMM-HMM 系统。

(4)用 MPE 方法优化 BN 特征提取网络参数和 GMM-HMM 模型参数,这也就是本文的创新所在。类似于 fMPE 训练,每次迭代都需要过 3 遍训练数据:第 1 遍计算 MPE 统计量;第 2 遍计算 fMPE 统计量,根据直接偏导和间接偏导计算目标函数对 BN 网络的梯度,然后更新 BN 特征提取网络参数;第 3 遍使用 Single pass retrain 更新 GMM 模型参数。

(5)用步骤(4)优化得到的网络来重新提取 BN 特征,并进一步使用模型空间 MPE 准则来优化 GMM-HMM 参数。

步骤(2)基于交叉熵准则训练 Bottleneck 神经网络目的是使交叉熵最小,即让 Tri-phone 状态之间的区分性更大。使用区分性准则来优化网络参数,目的是使区分性的目标函数最大,即让 Tri-phone 音素之间的区分性更大。

2.2 区分性目标函数

本文使用的目标函数是 BN 特征 $\boldsymbol{y}_r$ 和正确词序列 $\boldsymbol{w}_r$ 之间的最小音素错误率[4]

$$F_{\mathrm{MPE}}(\lambda)=\sum_r\sum_{\boldsymbol{w}_i\in\boldsymbol{w}^r}p(\boldsymbol{w}_i\mid\boldsymbol{y}_r)A(\boldsymbol{w}_i,\boldsymbol{w}_r)\approx\sum_r\sum_{\boldsymbol{w}_i\in\boldsymbol{w}^r}\frac{p_\lambda(\boldsymbol{y}_r\mid\boldsymbol{w}_i)P(\boldsymbol{w}_i)}{\sum_{\boldsymbol{w}_k\in\boldsymbol{w}^r}p_\lambda(\boldsymbol{y}_r\mid\boldsymbol{w}_k)P(\boldsymbol{w}_k)}A(\boldsymbol{w}_i,\boldsymbol{w}_r) \tag{3}$$

式中:$p(\boldsymbol{w}_i\mid\boldsymbol{y}_r)$为观察向量 $\boldsymbol{y}_r$ 对词序列 $\boldsymbol{w}_r$ 的后验概率;λ 为模型参数;$\boldsymbol{y}_r$ 为第 r 句 BN 特征;函数

$A(\boldsymbol{w}_i, \boldsymbol{w}_r)$为 $\boldsymbol{w}_i$ 和 $\boldsymbol{w}_r$ 之间的“净音素正确率”，基于 MPE 的区分性训练就是要最大化式(3)。

由于式(3)难以直接优化，一般通过构建弱辅助函数来优化

$$g_{\mathrm{MPE}}(\lambda, \bar{\lambda}) = \sum_r \sum_{q \in W_{\mathrm{lat}}^r} \sum_{t=s_q}^{t=e_q} \sum_m \gamma_q^{r\mathrm{MPE}} \gamma_{qm}^r(t) \log N(o_r(t), \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{4}$$

式中：$\gamma_q^{r\mathrm{MPE}}$ 为第 r 句训练语料中音素 q 以最小音素错误估测的概率，称为最小音素错误的权重；$\gamma_{qm}^r(t)$为弧 q 中第 m 个高斯在 t 时刻的高斯占有率。

2.3 更新网络最后一层

在这种策略下，网络中只有参数 $\boldsymbol{W}_3$(图 2 中连接 BN 层的权值)会在训练过程中更新。声学特征通过 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 映射到 $\boldsymbol{h}_2$，然后 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 保持不变，将 $\boldsymbol{h}_2$ 作为 $\boldsymbol{W}_3$ 的输入特征。由于 $y = \boldsymbol{W}_3 \boldsymbol{h}_2$，可以将 BN 特征提取网络中的 $\boldsymbol{W}_3$ 当做一个线性变换矩阵 $\boldsymbol{M}$，然后根据 MPE 准则来训练该矩阵，以得到更具有区分性的 BN 特征。这种训练方法与 fMPE[5] 相似，与 fMPE 不同的是，这里不需要使用 GMM 计算高维概率特征，也不需要构建一个新的高维矩阵来变换特征；这些都能用现有的 BN 特征提取网络实现，可以提高训练和解码速度。

矩阵 $\boldsymbol{M}$ 通过梯度算法来更新

$$M_{ij} := M_{ij} + \eta \frac{\partial F}{\partial M_{ij}} \tag{5}$$

式中：η 为学习率；$\dfrac{\partial F}{\partial M_{ij}}$为目标函数对 M_{ij} 的导数。

$$\frac{\partial F}{\partial M_{ij}} = \sum_{t=1}^{T} \frac{\partial F}{\partial y_{ti}} h_{tj} \tag{6}$$

式中：$\dfrac{\partial F}{\partial y_{t,i}}$表示目标函数对 t 时刻 BN 特征的第 i 维的导数；T 为全部训练集数据的帧数。使用大 Batch 代替 Mini-batch 更新网络参数，便于并行计算。

由于 GMM 模型参数 λ(主要是均值、方差)由变换之后的 BN 特征估计，目标函数对 BN 特征的导数包括直接偏导和间接偏导两部分。

$$\frac{\partial g_{\mathrm{MPE}}(\boldsymbol{M}, \bar{\boldsymbol{M}})}{\partial y_{t,i}} = \frac{\partial g_{\mathrm{MPE}}(y_{t,i}, \bar{\boldsymbol{M}})}{\partial y_{t,i}} + \frac{\partial g_{\mathrm{MPE}}(y_{t,i}, \bar{\boldsymbol{M}})}{\partial \lambda} \frac{\partial \lambda}{\partial y_{t,i}} \tag{7}$$

直接偏导为

$$\frac{\partial g_{\mathrm{MPE}}(y_{t,i}, \bar{\boldsymbol{M}})}{\partial y_{t,i}} = \sum_{q \in W_{\mathrm{lat}}^r} \sum_m \gamma_q^{\mathrm{MPE}} \gamma_{qm}(t) \frac{u_{m,i} - y_{t,i}}{\sigma_{m,i}^2} \tag{8}$$

式中：$u_{m,i}, \sigma_{m,i}$分别为第 m 个高斯的第 i 维的均值和方差，它们由 BN 特征根据最大似然准则估计得到。间接梯度由目标函数分别对 $u_{m,i}, \sigma_{m,i}$求导得到

$$\frac{\partial F(y_{t,i}, \bar{\boldsymbol{M}})}{\partial \lambda} \frac{\partial \lambda}{\partial y_{t,i}} = \sum_m \frac{\partial g_{\mathrm{MPE}}(\boldsymbol{M}, \bar{\boldsymbol{M}})}{\partial \tilde{\mu}_{m,i}} \frac{\partial \tilde{u}_{m,i}}{\partial y_{t,i}} + \sum_m \frac{\partial g_{\mathrm{MPE}}(\boldsymbol{M}, \bar{\boldsymbol{M}})}{\partial \tilde{\sigma}_{m,i}^2} \frac{\partial \tilde{\sigma}_{m,i}^2}{\partial y_{t,i}} \tag{9}$$

间接偏导的计算方法请参考文献[5]。

2.4 更新全部网络参数

在更新网络最后一层的策略下，由于最后一层网络的参数数量有限，可以通过求目标函数对 BN 特征提取网络的全部参数的梯度来更新整个网络。这种方法的训练过程与文献[7]提出的 NN-fMMI 方法类似，然而本文的目的是优化 BN 特征提取网络，而不是构建一个新的神经网络来变换 BN 特征。

在得到目标函数对 BN 特征的导数之后，更低层的网络参数的导数能够通过反向传播算法来计算。

$$\frac{\partial F}{\partial \boldsymbol{v}_3} = \frac{\partial F}{\partial \boldsymbol{y}_t}$$

$$\begin{aligned}
&\frac{\partial F}{\partial \boldsymbol{W}_3}=\frac{\partial F}{\partial \boldsymbol{v}_3}[\boldsymbol{h}_2^{\mathrm{T}};1] \quad \frac{\partial F}{\partial \boldsymbol{h}_2}=\boldsymbol{W}_3^{\mathrm{T}}\frac{\partial F}{\partial \boldsymbol{v}_3}\\
&\frac{\partial F}{\partial \boldsymbol{v}_2}=\sigma(\boldsymbol{v}_2)*(1-\sigma(\boldsymbol{v}_2))*\frac{\partial F}{\partial \boldsymbol{h}_2}\\
&\frac{\partial F}{\partial \boldsymbol{W}_2}=\frac{\partial F}{\partial \boldsymbol{v}_2}[\boldsymbol{h}_1^{\mathrm{T}};1] \quad \frac{\partial F}{\partial \boldsymbol{h}_1}=\boldsymbol{W}_2^{\mathrm{T}}\frac{\partial F}{\partial \boldsymbol{v}_2}\\
&\frac{\partial F}{\partial \boldsymbol{v}_1}=\sigma(\boldsymbol{v}_1)*(1-\sigma(\boldsymbol{v}_1))*\frac{\partial F}{\partial \boldsymbol{h}_1}\\
&\frac{\partial F}{\partial \boldsymbol{W}_1}=\frac{\partial F}{\partial \boldsymbol{v}_1}[\boldsymbol{x}^{\mathrm{T}};1]
\end{aligned} \tag{10}$$

式中:* 表示向量中各元素按顺序做相乘运算。网络参数 $\boldsymbol{W}$ 通过以下更新

$$\begin{aligned}
&\Delta \boldsymbol{W}_j^{(k+1)}=\alpha\Delta \boldsymbol{W}_j^{(k)}+\eta\frac{\partial F}{\partial \boldsymbol{W}_j}\\
&\boldsymbol{W}_j^{(k+1)}=\boldsymbol{W}_j^{(k)}+\Delta \boldsymbol{W}_j^{(k+1)} \quad j=1,\cdots,3
\end{aligned} \tag{11}$$

式中:α 为动量因子;η 为学习率。

3 实验配置及结果

实验建立在 Switchboard 任务上。训练集数据是从 Switchboard 语音库中随机抽取的 80 h 电话语音,采样率为 8 kHz。实验使用 NIST 2000 Hub5 测试集中的 Switchboard 部分,共 1 831 句话来测试语音识别性能。

3.1 基于交叉熵准则的 BN GMM-HMM 系统

实验采用的声学特征是 39 维 PLP 特征(能量和 12 维静态参数,以及它们的一阶差分和二阶差分),特征按说话人进行均值方差规整;BN 神经网络的输入由 11 帧(当前帧以及左右相邻各 5 帧)PLP 特征组成。输出包括 3 004 个 Tri-phone 绑定状态,每个状态 35 个高斯。语言模型用全部训练数据训练 3-gram 统计语言模型,根据第 2.1 节中步骤 1 至步骤 3 的描述训练基于 BN 特征的基线的 GMM-HMM 模型,表 1 是基线系统的词错误率(Word error rate, WER)。

表 1 基线系统词错误率 %

Tab. 1 Word error rate of baseline system

系统	MLE	MPE
基线系统	23.1	20.4

3.2 基于区分性准则更新 BN 网络最后层

为了测试间接梯度对实验性能的影响,本文进行了不更新模型和更新模型的对比实验,如表 2 所示。在本实验中,假设 GMM 的参数保持不变,只计算式(7)的直接梯度部分,这种方法和文献[16]描述的方法类似,不同的是这里使用 Batch 模式更新 BN 特征提取网络的最后一层,而不是使用 Mini-batch 的方式更新整个 BN 网络;表 2 第 1 行是使用这种方式更新的实验结果。表 2 中括号的数字表示直接使用不更新的 GMM-HMM 模型和优化后的 BN 特征进行解码得到的词错误率。使用优化后的 BN 特征重新根据 MLE 准则训练 GMM-HMM 模型将会使 WER 稍微增加,但是再次使用 MPE 准则训练得到的最终识别性能会更好,这种现象与文献[16]的实验描述一致。从实验结果来看,如果不更新模型,和基线系统对比 WER 下降很小,WER 绝对值下降只有 0.5%。在该实验中,将计算式(7)的直接偏导和间接偏导,每次迭代

表 2 只更新 BN 网络最后一层词错误率

Tab. 2 Word error rate of updating the last layer of BN neural networks %

模型	MLE	MPE
不更新模型	22.6 (21.9)	19.9
更新模型	21.4	19.1

GMM 模型参数和 BN 特征会同时更新;表 2 第 2 行是更新模型的实验性能。其中 MLE 对应 2.1 节中按步骤(4)更新后的模型,MPE 对应按步骤(5)更新后的模型。

实验结果表明,相对表 1,同时更新 GMM 模型和 BN 特征使 WER 下降更明显。与基线 BN 系统比较,MPE 之后的 WER 绝对下降了 1.3%,相对下降了 6.4%。更新 GMM 模型的性能明显优于不更新 GMM 模型的性能,说明式(7)中的间接梯度是很重要的部分。

图 3 是迭代过程中目标函数增长曲线。本文使用 Lattice 中总正确词个数除以总词数来反应目标函数增长;从图 3 中可以看出,更新模型比不更新模型的目标函数明显增长更多。

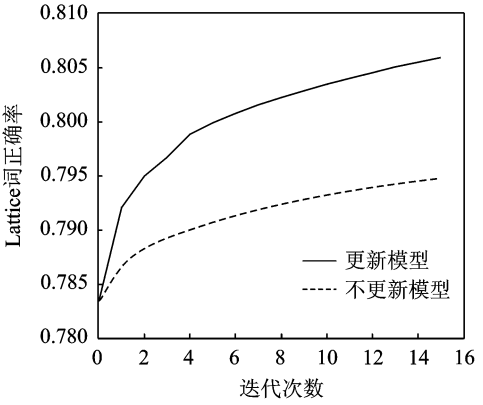

图 3 更新模型和不更新模型的目标函数变化曲线

Fig. 3 Objective function curve of updating and non-updating model

3.3 基于区分性准则更新全部 BN 网络

本实验将整个 BN 特征提取网络作为非线性特征变换,然后使用 BP 算法更新整个网络参数。实验使用了两种不同的方法来更新网络:第 1 种方法每次迭代先用 Batch 模式计算全部网络参数的梯度,然后同时更新参数;第 2 种方法从后向前逐层更新 BN 网络参数,每迭代更新完一层网络就将其参数固定,然后迭代更新前一层。表 3 为两种不同更新方式的实验性能。

表 3 更新全部 BN 网络词错误率

Tab. 3 Word error rate of updating the whole BN neural networks %

更新方式	MLE	MPE
同时更新	20.6	18.9
逐层更新	20.7	18.5

由表 3 可以看出,MPE 之后逐层更新方式比同时更新方式性能更好。优化之后的最小 WER 是 18.5%,比基线系统相对下降了 9.3%。同时,与表 2 比较,更新全部网络比只更新最后一层性能更好。

4 结束语

本文提出了一种基于格的 Bottleneck 特征提取网络区分性训练方法,使用了两种不同的方法来构建特征空间变换。第 1 种方法将网络最后一层参数当做一个线性变换,计算更加简便快捷;第 2 种方法将 BN 特征提取网络看做一个非线性变换,更新整个网络的参数。通过采用最小音素错误率准则,在训练过程中同时优化 BN 特征提取网络和 GMM 参数。本文使用所有训练数据来计算目标函数对网络的梯度,适合分布式计算,可以加快训练速度。实验结果表明,相对传统的 BN 特征提取网络,本文提出的两种方法都能提升语音识别率,其中第 1 种方法词错误率相对基线下降 6.4%,第 2 种方法词错误率相对基线下降 9%。

参考文献:

[1] Kapadia S, Valtchev V, Young S J. MMI training for continuous phoneme recognition on the TIMIT database[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Minnesota, USA: IEEE, 1993:491-494.

[2] Juang B H, Chou W, Lee C H. Minimum classification error rate methods for speech recognition[J]. IEEE Transactions, 1997,5(3): 257-265.

[3] Mc Dermott E, Hazen T, Roux J L, et al. Discriminative training for large vocabulary speech recognition using minimum classification error[J]. IEEE Transactions, 2007,15(1):203-223.

[4] Povey D, Woodland P. Minimum phone error and I-smoothing for improved discriminative training[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Florida, USA: IEEE, 2002:105-108.

[5] Povey D, Kingsbury B, Mangu L, et al. FMPE: Discriminatively trained features for speech recognition[C]//Proceedings of

International Conference on Acoustics, Speech and Signal Processing. Philadelphia, USA: IEEE, 2005:961-964.
[6] Povey D, Kanevsky D, Kingsbury B, et al. Boosted MMI for model and feature-space discriminative training[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008:4057-4060.
[7] Saon G, Kingsbury B. Discriminative feature-space transforms using deep neural networks[C]//Proceedings of International Speech Communication Association. Portland, USA: IEEE, 2012.
[8] 余华,黄程韦,金赟,等.基于粒子群优化神经网络的语音情感识别[J].数据采集与处理,2011,26(1):57-62.
Yu Hua, Huang Chengwei, Jin Yun, et al. Speech emotion recognition based on particle swarm optimizer neural network [J]. Journal of Data Acquisition and Processing, 2011,26(1):57-62.
[9] 徐以中.神经网络模拟实验与语言认知研究的互动[J].南京航空航天大学学报,2010,12(1):75-79.
Xu Yizhong. Interaction between artificial neural network computation and cognitive perspective on language study[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2010,12(1):75-79.
[10] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition [J]. IEEE Transactions, 2012,20(1):30-42.
[11] Kingsbury B. Lattice-based optimization of sequence classification criteria for neural network acoustic modeling[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Taipei, China: IEEE, 2009:3761-3764.
[12] Bridle J S, Dodd L. An Alphanet approach to optimizing input transformations for continuous speech recognition[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE, 1991:277-280.
[13] Grezl F, Karafiat M, Kontar S, et al. Probabilistic and bottle-neck features for LVCSR of meetings[C]//Proceedings of International Conference on Acoustics, Speech and Signal Processing. Hawaii, USA: IEEE, 2007:757-760.
[14] Hu H B, Zahorian S A. A neural network based nonlinear feature transformation for speech recognition[C]//Proceedings of International Speech Communication Association. Brisbane, Australia: IEEE, 2008:1533-1536.
[15] Yu D, Seltzer M L. Improved bottleneck features using pretrained deep neural networks[C]//Proceedings of International Speech Communication Association. Florence, Italy: IEEE, 2011: 237-240.
[16] Paulik M. Lattice-based training of bottleneck feature extraction neural networks[C]//Proceedings of International Speech Communication Association. Lyon, France: IEEE, 2013:89-93.

作者简介:

刘迪源(1990-),男,硕士研究生,研究方向:语音识别,E-mail: ldy2012 @ mail.ustc.edu.cn。

郭武(1973-),男,副教授,研究方向:语音识别、说话人识别、语种识别和音频检索。