文章编号：2095-2163(2020)12-0001-06 中图分类号：TP391.41 文献标志码：A

基于快速近似时序池化的端到端声学事件识别

张力文，韩纪庆
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：声学事件识别系统的性能很大程度上取决于音频特征学习的有效性。由于音频信号属于时序性信号，要获得有效的音频特征，就需要提取其中的时序信息。作者曾提出了一种有效的时序性特征学习方法：时序池化。然而，由于其需要求解一个没有闭式解的优化问题，导致无法灵活地运用在当前流行的深度学习框架之中。为此，本文在保留时序池化的前提下，提出了一种计算方式更为简单的快速近似时序池化方法。基于此方法又进一步提出一种用于解决端到端声学事件识别问题的卷积神经网络。实验结果表明，所提出的网络可以取得比目前大多数方法更好的识别性能。

关键词：声学事件识别；音频特征学习；时序池化；卷积神经网络

End-to-end acoustic event recognition with fast approximate temporal pooling

ZHANG Liwen, HAN Jiqing
(School of Computer Science and Technology, Harbin Institude Of Technology, Harbin 150001, China)

【Abstract】 The performance of an acoustic event recognition (AER) system depends largely on the effectiveness of audio representation learning. Since audio signal is temporally-structured, to obtain effective audio representation, it is necessary to extract the temporal information. In our previous work, we had proposed an effective method for learning temporal features, i.e., temporal pooling. However, this method involves solving an optimization problem without closed-form solution, which limits its flexibility in being performed as an intermediate module of the deep learning frameworks. In this paper, we propose a fast approximate calculation method for the temporal pooling. Furthermore, based on this approximate method, we propose an end-to-end convolutional neural network for the AER task, which can outperform most of the state-of-the-art AER systems.

【Key words】 Acoustic Event Recognition; Audio Representation Learning; Temporal Pooling; Convolutional Neural Network

0 引 言

声学事件指的是由某个物理声源产生的声响活动，如车辆经过、鸟叫和门铃声等[1]。声学事件识别(Acoustic Event Recognition, AER)的目标是利用音频信号处理及机器学习等技术手段，对声音信号中的内容进行分析和提取，从而判断或识别出其所属的事件类别[2]。AER 作为视听觉感知计算的重要组成部分之一，在听觉相关的智能人机交互，以及海量音视频多媒体数据处理等领域有着广阔的应用前景，如语音分离和增强技术[3]、基于内容的音频检索[4-5]，以及安防监控领[6]域等，还能用于提升移动可穿戴计算设备的用户体验[7]。

声学事件识别通常包含特征表示提取或学习和分类器训练两个部分，其中特征表示的有效性对于系统识别性能的好坏至关重要。早期的特征提取方法大都通过计算音频样本帧级别声学特征，如基频、过零率、子带能量及梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficient，MFCC)等的均值和方差等来获得样本的特征表示[4-5,8-9]。然后使用基于最近邻(Nearest Neighbor, NN)准则的分类方法[4-5\]或支持向量机(Support Vector Machine, SVM)[8-9]来对样本的事件类别进行识别判断。然而完整的声学事件通常都会持续数秒甚至更长的时间，仅包含几十毫秒(20~50 ms)声音内容的帧级别特征中可供分析的上下文内容极为有限；基于统计的特征提取方式也会破坏音频信号中原有的时频结构，从而造成大量有用信息的丢失，不利于得到有效的音频特征表示。鉴于此，为在特征学习过程中利用更多的上下文信息，越来越多研究工作试图在较长时的音频片段上提取出特征表示，例如利用混合高斯模型(Gaussian Mixture Model, GMM)来刻画音频片段中连续多帧 MFCC 特征的分布[10]；或者利用音频词袋(Bag-of-Audio-Words, BoAW)方法，统计连续多个采用基于聚类的字典学习方法得到的帧级

基金项目：国家重点研发项目(2017YFB1002102)；国家自然科学基金(91120303)。

作者简介：张力文(1991-)，男，博士研究生，主要研究方向：音频信号处理、声学事件及场景识别与分类；韩纪庆(1964-)，男，博士，教授，博士生导师，主要研究方向：语音信号处理、音频信息处理。

收稿日期：2020-11-01

编码结果序列的直方图[11]。随着近十年来深度学习技术在音频信号处理各研究领域的成功应用,其已经成为声学事件识别的主流方法[12-13]。与之前的方法不同,基于深度学习的 AER 普遍是以端到端的方式对音频特征和分类器进行联合学习的,不需要单独训练分类器。其中,具有代表性的工作有使用音频片段中多个帧级别特征拼接而成的长向量训练深度神经网络(Deep Neural Network, DNN),以同时得到片段级特征表示和分类器参数,最终对样本内多个片段的识别结果投票以产生样本的识别结果[14]。借鉴由牛津大学视觉几何学研究组(Visual Geometry Group, VGG)提出的 VGG [15]卷积神经网(Convolutional Neural Network, CNN)的架构,而设计的声学事件识别网络 AENe[16];以及通过利用包含 128 个 1×1 共享参数卷积核的卷积层和更大步长的卷积层,来分别替换 AENet 中的全连接层和 max-pooling 层,从而获得网络参数更少、识别性能更好的 CNN-C[17]。与 BoAW、GMM 和 DNN 等片段级特征提取和学习方法相比较,CNN 可以保证所学特征对于输入具有平移不变(shift-invariant),因而能够有效地保持输入片段的时频结构,同时堆叠的卷积操作能从多种时频尺度上提取输入片段的上下文内容。此外,卷积核的参数共享机制也有助于其能以比 DNN 更少的网络参数,对更长时的输入进行学习。

以上工作对声学事件识别的研究起到了极大的推动作用,尤其是在特征提取及学习方面,但这些方法仍然存在一定的局限性。考虑到声音信号本质上是时序性的信号,要学习出能够刻画音频样本完整语义内容的特征表示,就需要获取音频样本之中的时序信息。而之前提到的方法在特征提取或学习阶段,都未明确考虑音频帧与帧,以及片段与片段之间的时序依赖关系,因而无法获得具有完整时序信息的音频特征表示。为此,目前有研究工作将循环神经网络(Recurrent Neural Network, RNN)及其变种等基于时间序列建模的方法,引入到音频特征学习阶段,试图进一步提升 AER 的性能。例如,Meyer 等将 CNN 与基于长短时记忆(Long Short-Term Memory, LSTM)的 RNN 网络相结合,并提出一种音频特征自编码网络 ConvLSTM[18];Sang 等提出的面向端到端环境音分类的卷积循环网络(Convolutional RNN, CRNN)[19]。RNN 等方法的引入使得网络可以在特征学习阶段,捕捉到局部特征表示间的时序关系,然而 RNN 及其变种网络的长时遗忘效应,导致它们对较长序列中完整时序信息的刻画能力存在着一定的局限性,因而对 AER 系统性能的提升也很有限。为探索更有效捕获时序信息的特征学习方法,作者在之前的研究工作中提出了一种基于支持向量回归(Support Vector Regression, SVR)[20]的无监督特征学习方法——时序池化(Temporal Pooling, TP)[21],以及其有监督的形式,即基于可学习区分性映射的时序池化(Discriminative Mapping Learned TP, DM-TP)[22]。利用预训练好的 CNN 得到的音频样本的片段级特征序列作为输入,DM-TP 捕捉片段之间的时序依赖关系,并同时更新分类器参数,以实现时序关系依赖的音频特征及分类器的联合学习。

然而,DM-TP 的训练过程实际上是求解一个以 SVR 优化问题为约束的双层优化问题[23]。由于该求解过程需要耗费较大计算资源,导致其无法像一般的最大池化或平均池化操作一样,灵活地嵌入至一般的神经网络之中高效地端到端训练,通常需要单独训练片段级的局部特征学习 CNN 网络,以构建出 DM-TP 模型的输入序列。为此,本文从简化 TP 中时序编码问题的梯度计算入手,提出一种近似求解 TP 时序编码问题的方法,称之为快速近似时序池化(Fast Approximate TP, FATP)。通过将基于 SVR 的时序编码问题转化为具有闭式解的时序编码函数,FATP 可以在保证不损失过多时序编码性能的情况下,灵活地插入至各种经典的 CNN 网络中,以实现局部特征、时序信息及分类器的联合学习。基于这一优势,进一步提出了端到端的声学事件识别网络 FasTemP-Net。

1 基于 SVR 的时序池化

为更好地解释如何实现 FATP,以及在端到端网络的框架中研究快速计算 TP 方法的必要性,首先简要介绍 TP 方法和 DM-TP 方法的输入特征序列构建过程:

(1)片段级特征提取:给定一个预训练好的片段级特征提取 CNN 网络 $\Psi_{CNN}(\cdot;\boldsymbol{W})$,其中 $\boldsymbol{W}$ 为网络参数;对于某一音频样本的频谱片段序列 $\{\boldsymbol{X}_1^0,\ldots,\boldsymbol{X}_T^0\}$ 中的每一个 $\boldsymbol{X}_t^0(t=1,\ldots,T)$,CNN 特征提取网络将用于得到其特征表示 $\Psi(\boldsymbol{X}^0{}_t;\boldsymbol{W})=x_t \in \mathbb{R}^C$。相应地,可得整个样本的片段级特征表示序列 $\boldsymbol{X}=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T] \in \mathbb{R}^{C\times T}$,其中 C 是特征维度。

(2)特征序列预处理:一个逐点操作的非线性特征映射算子 $\Psi(\cdot):\mathbb{R}^C \rightarrow \mathbb{R}^P$ 和时移平均平滑(Time Varying Mean, TVM)操作将共同作用于 $\boldsymbol{X}$,以得到 TP 的输入特征序列 $\boldsymbol{V}=[v_1,\ldots,v_T] \in \mathbb{R}^{P\times T}$,其中每个 v_t 的计算方式如式(1):

$$\boldsymbol{v}_t = \sum_{\tau=1}^{t} \Psi(\boldsymbol{x}_\tau),\ t = 1,\dots,T. \tag{1}$$

其中，若 $\Psi(\cdot)$ 采用 posneg 核函数，则 $\boldsymbol{v}_t$ 的维度 $P = 2C$。

在得到输入特征序列 $\boldsymbol{V}$ 之后，TP 的目的是对 $\boldsymbol{V}$ 中的时序动态信息 D 进行编码，以得到序列的时序性特征表示。其大致思想为利用一个参数可学习的线性函数 $f(\boldsymbol{V};\ u) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{V}$ 来拟合 D，并用参数 $\boldsymbol{u}$ 来作为 $\boldsymbol{V}$ 的时序性特征表示，该参数的学习问题可近似表示为形式(2)：

$$\arg\min_{u} \| D - f(V;\ u) \|. \tag{2}$$

由于 D 反映的是 $\boldsymbol{V}$ 中各时刻特征向量在时间方向上的变化趋势，即每一个向量是如何按照 $\boldsymbol{V}$ 原本的时间顺序依次出现的，因此，$f(\cdot;\boldsymbol{u})$ 还需要满足以下约束条件(3)：

$$\forall t_i, t_j \in \{1,\dots,T\},\ t_i < t_j \Leftrightarrow f(v_{t_i};\ u) < f(\boldsymbol{v}_{t_j};\ \boldsymbol{u}). \tag{3}$$

为在这一约束下，实现 $\boldsymbol{u}$ 的学习，TP 采用一种基于回归的逐元素优化策略，即将 $\boldsymbol{v}_t$ 作为回归因子，其时间索引 t 作为标签，通过求解式(4)所示的正则化 SVR 问题来得到最优参数 $\boldsymbol{u}^* \in \mathbb{R}^P$：

$$\boldsymbol{\Phi}(\boldsymbol{V}) = \boldsymbol{u}^*\ a\ \arg\min_{u} E(\boldsymbol{V};\ \boldsymbol{u}),$$
$$E(\boldsymbol{V};\ \boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{u}\|^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\left[\,|t - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}_t| - \varepsilon\right]^2 \geqslant 0. \tag{4}$$

其中，$E(\cdot;\boldsymbol{u})$ 为 SVR 优化问题的目标函数；$[\cdot]_{\geqslant 0} = \max\{\cdot, 0\}$ 为 ε - 不敏感损失函数；λ 为正则项系数。通过求解式(4)所示的优化问题，则可得到 $\boldsymbol{V}$ 的时序性特征表示。

而在有监督的 DM-TP 中，式(4)所示的优化问题将以底层约束条件的形式出现在如式(5)所示的双层优化问题中：

$$\begin{aligned} &\min_{M,\theta} L(\boldsymbol{u}^*,\ y;\ \boldsymbol{\theta}) + R_\theta + R_M, \\ &\text{s.t. } \boldsymbol{u}^* \triangleq \arg\min_{u} E(\boldsymbol{X};\ \boldsymbol{M}, \boldsymbol{u}). \end{aligned} \tag{5}$$

其中，$L(\boldsymbol{u}^*,\ y;\ \boldsymbol{\theta})$ 为顶层分类器的目标函数；$\boldsymbol{\theta}$ 为分类器参数，若采用 softmax，则为交叉熵损失函数；y 为片段级特征序列 $\boldsymbol{X}$ 对应样本的类别标签；R_θ 与 R_M 分别为关于分类器参数 θ 和区分性映射(Discriminative Mapping，DM)权重矩阵 $\boldsymbol{M} \in \mathbb{R}^{C\times C}$ 的正则项。值得注意的是，这里 DM 与 TP 中的非线性特征映射算子不同，其是一个参数可学习的特征映射 $\boldsymbol{\Psi}_{DM}(\cdot;\boldsymbol{M}):\mathbb{R}^C \to \mathbb{R}^C$，其将取代 TP 中的预处理阶段，以将顶层分类器中学到的类别信息反馈至底层时序编码过程之中。其具体形式如式(6)：

$$\boldsymbol{v}_t = \psi(\boldsymbol{M}\boldsymbol{x}_t),\ t = 1,\dots,T. \tag{6}$$

其中，ψ 为某个激活函数，例如 ReLU。

具体实施过程中，采用基于梯度的优化方法对 DM-TP 双层优化问题求解。由于底层 TP 问题本身是一个没有闭式解的优化问题，因此问题求解最困难的部分在于损失函数 $L(\cdot,\cdot;\boldsymbol{\theta})$ 关于 $\boldsymbol{M}$ 的梯度计算。若忽略正则项，$L(\cdot,\cdot;\boldsymbol{\theta})$ 关于 $\boldsymbol{M}$ 中的每一个元素 M_{ij} 的梯度计算表达式可推导为式(7)[22]：

$$\frac{\partial L}{\partial M_{ij}} = \left(\frac{\partial L}{\partial \boldsymbol{u}^*}\right)^{\mathrm{T}} \frac{\partial \boldsymbol{u}^*}{\partial M_{ij}}. \tag{7}$$

其中，

$$\begin{aligned} &\frac{\partial \boldsymbol{u}^*}{\partial M_{ij}} = -\boldsymbol{H}_u^{-1}\boldsymbol{H}_{M_{ij}}, \\ &\boldsymbol{H}_u = \frac{\partial^2 E}{\partial \boldsymbol{u}^* \partial \boldsymbol{u}^*} = I + \lambda \sum_{l_t \neq 0} \boldsymbol{v}_t \boldsymbol{v}_t^{\mathrm{T}}, \\ &\boldsymbol{H}_{M_{ij}} = \frac{\partial^2 E}{\partial M_{ij} \partial \boldsymbol{u}^*} = -\lambda \sum_{l_t \neq 0} \left(\frac{\partial \boldsymbol{v}_t}{\partial M_{ij}} + \boldsymbol{u}^{*\mathrm{T}} \frac{\partial \boldsymbol{v}_t}{\partial M_{ij}} \boldsymbol{v}_t \right). \end{aligned} \tag{8}$$

其中，l_t 为 $\boldsymbol{v}_t$ 在 $\boldsymbol{u}^*$ 下的 ε - 不敏感损失值。从式(8)可知，要求出 $\boldsymbol{M}$ 的梯度则需计算 Hessian 矩阵 $\boldsymbol{H}_u$ 的逆，其计算需要 $O(C^3)$ 的时间复杂度，当特征维度 C 较大，此环节将会耗费大量计算资源。若将 DM-TP 作为 CNN 网络中的某一个中间模块，则区分性映射将被堆叠式的卷积操作所取代，则模型将会更加复杂，梯度计算也会更加费时。因此，基于 DM-TP 的端到端 CNN 网络并不具备现实意义，需要一种更加快速的近似方法来对 TP 进行简化，使其能灵活地嵌入至 CNN 网络中。

2　基于快速近似时序池化的声学事件识别

由于 DM-TP 将 TP 中基于 SVR 的时序编码作为底层约束，因而导致其梯度计算过程过于复杂，不利于以中间层的形式出现在一般的 CNN 网络之中端到端的训练。为此，本文尝试将 TP 时序编码问题的求解过程进行简化，以使其能通过一个具有闭式解的函数来计算，进而可以像一般的池化操作一样能直接将编码结果代入到网络顶层的分类器目标损失函数中，并利用反向传播算法进行快速梯度计算。

2.1　快速近似时序池化

根据式(3)所示 TP 的约束条件可知，拟合函数 $f(\cdot;\boldsymbol{u})$ 可以看作是一种序列的排序打分函数，其参数就可看作是序列排序规则的表示。在之前的工作

采用基于逐元素优化的策略来对参数求解,而实际上,其也可采用如式(9)所示的两两成对的优化策略:

$$\begin{cases}\Phi(\boldsymbol{V}) = \boldsymbol{u}^* \triangleq \underset{u}{\operatorname{argmin}} E(\boldsymbol{V};\ \boldsymbol{u}),\\ E(\boldsymbol{V};\ \boldsymbol{u}) = \dfrac{1}{2}\|\boldsymbol{u}\|^2 + \lambda \sum_{r>t} \max\{0,\ 1 - \\ \qquad f(\boldsymbol{v}_r;\ \boldsymbol{u}) + f(\boldsymbol{v}_t;\ \boldsymbol{u})\}.\end{cases} \tag{9}$$

其中,目标函数 $E(\cdot;\boldsymbol{u})$ 中的第一项可以看成是 SVM 中的二次正则项,第二项则是 hinge-loss 损失函数,记录了那些不能被 $f(\cdot;\boldsymbol{u})$ 正确排序的 r 和 t 时刻的特征向量对的数量,从而保证目标函数可以对 $\forall r,t \in \{1,\ldots,T\}$, $r > t$, 满足约束条件 $f(\boldsymbol{v}_r;\ \boldsymbol{u}) > f(\boldsymbol{v}_t;\ \boldsymbol{u}) + 1$。

推导快速近似时序池化 FATP 的关键,在于如何对式(9)中目标函数关于 $\boldsymbol{u}$ 的梯度近似计算。首先从 $\boldsymbol{u} = \vec{0}$ 开始,式(9)所示优化问题第一次迭代后的近似解应为式(10):

$$\forall \eta > 0,\ \boldsymbol{u}^* = \vec{0} - \eta\, \tilde{\nabla} E(\boldsymbol{V};\ \boldsymbol{u})\big|_{\boldsymbol{u}=\vec{0}} \propto - \tilde{\nabla} E(V;\ u)\big|_{\boldsymbol{u}=\vec{0}}. \tag{10}$$

其中,η 为大于零的学习率,且 $\tilde{\nabla} E(\boldsymbol{V};\ \boldsymbol{u})\big|_{\boldsymbol{u}=\vec{0}}$ 可近似为式(11):

$$\tilde{\nabla} E(\boldsymbol{V};\ \boldsymbol{u})\big|_{\boldsymbol{u}=\vec{0}} \propto \sum_{r>t} \tilde{\nabla} \max\{0,\ 1 - f(\boldsymbol{v}_r;\ \boldsymbol{u}) + f(\boldsymbol{v}_t;\ \boldsymbol{u})\}\big|_{\boldsymbol{u}=\vec{0}} = \sum_{r>t} \tilde{\nabla} \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{v}_t - \boldsymbol{v}_r)\big|_{\boldsymbol{u}=\vec{0}} = \sum_{r>t}(\boldsymbol{v}_t - \boldsymbol{v}_r). \tag{11}$$

则,$\boldsymbol{u}^*$ 可有如式(12)近似的表达形式:

$$\boldsymbol{u}^* \propto \sum_{r>t}(\boldsymbol{v}_t - \boldsymbol{v}_r) = (\boldsymbol{v}_2 - \boldsymbol{v}_1) + (\boldsymbol{v}_3 - \boldsymbol{v}_2) + (\boldsymbol{v}_3 - \boldsymbol{v}_1) + \ldots + (\boldsymbol{v}_T - \boldsymbol{v}_{T-1}) + \ldots + (\boldsymbol{v}_T - \boldsymbol{v}_1) = \sum_{t=1}^{T} \sigma_t \boldsymbol{v}_t. \tag{12}$$

其中,σ_t 可看成是 $\boldsymbol{v}_t$ 的加权系数,推出其计算表达式(13):

$$\sigma_t = (t-1) - (T-t) = 2t - T - 1. \tag{13}$$

若考虑 TP 中的特征预处理,则式(12)可进一步写成式(14):

$$\boldsymbol{u}^* \propto \sum_{t=1}^{T} \sigma_t v_t = \sum_{t=1}^{T} \zeta_t \Psi(\boldsymbol{x}_t). \tag{14}$$

根据式(1)和(12),可以推出加权系数 ζ_t 的计算表达式(15):

$$\zeta_t = 2(T - t + 1) - (T + 1)(S_T - S_{t-1}). \tag{15}$$

其中,$S_t = \sum_{i=1}^{t} 1/t$ 为调和序列的前 t 项之和,且 $S_0 = 0$。

由于 FATP 在 CNN 网络中是用来学习各卷积模块得到的局部特征表示之间的时序关系,而卷积模块本身就可看作是非线性的特征映射,因此 FATP 无需再对输入特征进行特征映射,则式(14)中的 $\Psi(\cdot)$ 可直接采用等价映射,最终可得 FATP 的计算公式(16):

$$\hat{\Phi}(\boldsymbol{X}) = \sum_{t=1}^{T} \zeta_t \boldsymbol{x}_t. \tag{16}$$

2.2 基于 FATP 的 FasTemP-Net

为在实验过程中,公平地对比 FasTemP-Net 和之前所提出 DM-TP 之间的性能,公平地对比 FasTemP-Net 和之前所提出 DM-TP 之间的性能,采用 Event-Net[22] 作为所提出网络的骨干网络。图 1 给出了两种情况下 FasTemP-Net 的网络结构,网络的输入为某一音频样本按时间顺序排列的对数 Mel 域频谱片段序列。其中,第一种为直接使用单个全局 FATP 层的 FasTemP-Net,在此种情况下,FATP 将会对全连接层(Fully-Connected, FC)输出特征序列进行单一时间尺度上的时序信息学习;第二种情况是借鉴金字塔时序池化(Pyramidal Temporal Pooling, PTP)[22] 的多尺度学习思想,即通过将单个 FATP 层推广为多个堆叠式的局部 FATP 操作,来实现在多种时间尺度上对特征序列中的时序信息进行学习。

FasTemP-Net 的具体实现可分为以下两个步骤:(1)将预训练好的针对片段级特征提取的骨干网络 Event-Net 中各个卷积层(Conv)和 FC 层的参数设置为共享,然后将 FATP 插入相邻的 FC 层之间(如 FC5 和 FC6),以得到针对样本级声学事件识别的 FasTemP-Net;(2)对得到的 FasTemP-Net 微调训练,训练算法采用基于动量加速的随机梯度下降算法(Stochastic Gradient Decent, SGD)[24]。

3 实 验

实验采用包含 28 种声音事件的 AER 数据集[16],其总共包括 5 223 个音频样本,时长共计 768.4 min。每个类别的样本数从 59—378 不等,单个样本时长为 3-12 s,训练集与测试集分别占整个数据集的 75% 和 25%。所有样本均统一转换为采样率 16 kHz、精度 16 bits 及单声道的 wav 格式。然后提取样本的 Log-Mel 频谱作为前端特征,其中特征维度为 129,帧长和帧移分别为 40 ms 和 10 ms。FasTemP-Net 的训练参数的具体设置如下:初始学习率为 0.001,最大迭代次数为 20,动量因子为 0.9,权重衰减系数为 0.001,mini-batch 的大小为 32。微调训练过程中网络的学习率会随着迭代次数的增加而以对数级别下降,学习率的最小值设为 0.000 01。

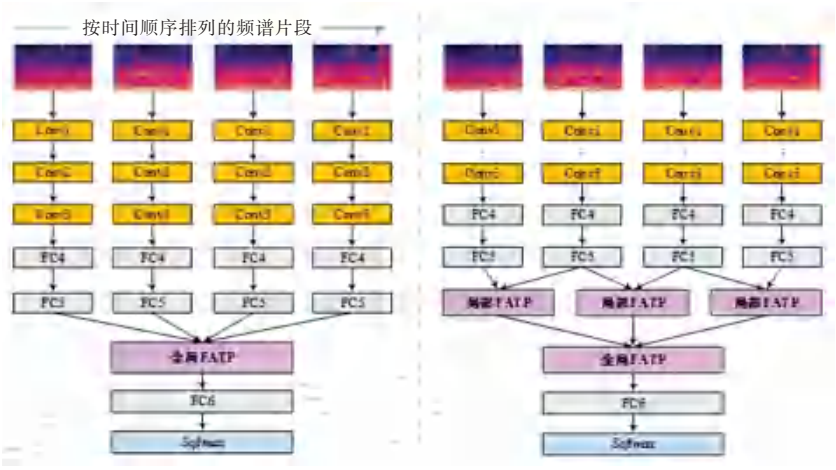

图 1　FasTemP-Net 网络结构示意图

Fig. 1　Architectures of FasTemP-Nets

为验证 FATP 的有效性，首先对比分别使用 FATP 与其它池化方法的 Event-Net 的识别性能，即在不进行端到端训练的情况下，观察 FATP 对网络性能的提升。对比结果见表 1。其中，Voting 的识别结果是通过对 Event-Net 输出的片段级识别结果投票而得到的；Avg/Max-Pooling 分别表示平均和最大池化；SVM 分类器采用卡方核函数。从表 1 可以看出，FATP 和 TP 由于考虑了时序信息均可比其他池化方法获得更好的识别性能，考虑了多尺度时序信息的 PTP 与 MS-FATP 还能进一步提升系统的性能；此外还发现简化后的 FATP 可以获得与 TP 相近的识别性能，在实验过程中通过统计二者的处理速度发现，FATP 的运算速度比 TP 快了将近 50 倍。

表 1　FATP 与其他池化方法的性能对比

Tab. 1　Performance comparison of FATP and other pooling methods

方法	识别正确率/%
Event-Net + Voting	82.1
Event-Net + Avg-Pooling + SVM	81.3
Event-Net + Max-Pooling + SVM	81.0
Event-Net + TP + SVM	84.5
Event-Net + PTP + SVM	87.5
Event-Net + FATP + SVM	84.1
Event-Net + MS-FATP + SVM	86.9

最后，对基于 FATP 的 FasTemP-Net 实验验证，以观察在端到端训练的情况下，FATP 对网络性能的提升。单尺度 FasTemP-Net 和多尺度 MS-FasTemP-Net 的识别性能表现见表 2，对比了其与目前其它主流方法的识别性能。从结果可看出：

(1)与 AENet[22] 和 CNN-C[23] 等基于经典 VGG 网络架构的 CNN 相比，考虑了局部特征间时序关系的 FasTemP-Net 明显更具性能优势；

(2)与无监督的时序性特征学习方法 TP[21] 相比，由于所提出的 FasTemP-Net 可在时序池化阶段利用分类器学到的类别信息，因此也能取得更好的识别性能；

(3)与同样基于多尺度时序信息学习的 PTP 和 DM-PTP[22] 相比，由于 FATP 的引入，使得 MS-FasTemP 能以更简洁的形式进行高效的端到端，并且不会因为使用了更为简化的时序池化操作，而带来明显的性能下降。

表 2　FasTemP-Net 与其他主流方法的性能对比

Tab. 2　Performance comparison of FasTemP-Net and state-of-the-art methods

方法	识别正确率/%
AENet[22]	80.3
CNN-C[23]	86.0
ConvLSTM + softmax[25]	78.9
BoAW_SC + TP + SVM[29]	84.4
Event-Net + PTP + SVM[30]	88.9
Event-Net + DM-PTP[30]	89.9
FasTemP-Net	86.9
MS-FasTemP-Net	89.0

4　结束语

时序性特征学习方法 TP 可以有效的获取音频样本中的时序信息，由于其涉及 SVR 问题求解，导致无法灵活的嵌入到当前流行的深度神经网络之中。针对这一问题，本文给出了 TP 的快速近似计算方法 FATP。进一步提出了基于 FATP 的端到端声学事件识别网络 FasTemP-Net。实验结果显示，所提出的网络可以在不过多损失时序信息学习能力的同时，取得比目前大多数方法更好或者近似的识别性能。

(下转第 11 页)