计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(11):147−152 [doi: 10.15888/j.cnki.csa.007143]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于迁移学习的暴恐音频判别方法①

胡鑫旭, 周　欣, 何小海, 熊淑华, 王正勇

(四川大学 电子信息学院, 成都 610065)
通讯作者: 何小海, E-mail: hxh@scu.edu.cn

摘　要: 本文从网络和电影中截取暴恐音频片段组成暴恐音频库, 由于暴恐音频来源受限, 而卷积神经网络需要大量的数据训练, 为此, 将迁移学习技术引入暴恐音频的判别中. 首先采用公开的 TUT 音频数据集进行预训练, 然后保留模型权重并迁移网络在暴恐音频库上继续训练, 最后在 fine-tune 后的网络中增加网络的层数, 添加了一种类似于残差网络的结构使其能够利用更多的音频信息. 实验结果表明, 使用迁移学习方法比未使用迁移学习方法的平均判别率提升了 3.97%, 有效解决了在暴恐音频判别研究中音频数据集过小而带来的训练问题, 且改进后的迁移学习网络进一步提升了 1.01% 的平均判别率, 最终达到 96.97% 的判别率.

关键词: 暴恐音频判别; 迁移学习; 卷积神经网络; 深度学习; 残差网络

引用格式: 胡鑫旭,周欣,何小海,熊淑华,王正勇.基于迁移学习的暴恐音频判别方法.计算机系统应用,2019,28(11):147–152. http://www.c-s-a.org.cn/1003-3254/7143.html

Discrimination Method of Terrorism Audio Based on Transfer Learning

HU Xin-Xu, ZHOU Xin, HE Xiao-Hai, XIONG Shu-Hua, WANG Zheng-Yong

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: This article intercepts the horror audio clips from the network and movies to build terrorism audio dataset. However, the source of the horror audio is limited, whereas the convolutional neural network depends on a large amount of data. To this end, the transfer learning technology is performed in the discrimination of the terrorism audio. Firstly, pre-train the network by using the public TUT acoustic scenes dataset, and then retain the model weight and transfer the neural network to the discrimination of terrorism audio. Finally, add more layers after the fine-tune network to utilize more audio information, the structure of the added layers is similar to the residual network. The experimental results indicate that the average discriminant rate of the transfer learning method is 3.97% higher than that of the non-transfer learning method, which effectively solves the training problem caused by small audio dataset in the study of terrorism audio discrimination, and the average discriminant rate of the improved transfer learning network has increased by 1.01%, finally reaches the discriminant rate of 96.97%.

Key words: discrimination of terrorism audio; transfer learning; convolutional neural network; deep learning; residual network

① 基金项目: 国家自然科学基金 (61871278); 成都市产业集群协同创新项目 (2016-XT00-00015-GX); 四川省科技计划项目 (2018HH0143); 四川省教育厅科研项目 (18ZB0355)

Foundation item: National Natural Science Foundation of China (61871278); Industrial Cluster Collaborative Innovation Project of Chengdu Municipality (2016-XT00-00015-GX); Science and Technology Program of Sichuan Province (2018HH0143); Science and Technology Research Program of Education Bureau, Sichuan Province (18ZB0355)

收稿时间: 2019-04-11; 修改时间: 2019-05-08; 采用时间: 2019-05-13; csa 在线出版时间: 2019-11-06

随着近年来互联网与电影行业的快速发展, 网络上包含的音视频信息与日俱增, 为用户所共享的音视频中不乏包含有暴力恐怖音视频, 这些暴恐音视频将产生不良的网络环境, 对缺乏判断力的未成年人产生负面影响. 针对此现象, 通常由人工进行审查, 审核通过以后才可进入网络, 但由于网络上的音视频信息丰富, 并且每日还会产生海量的音视频, 所以这种做法不仅耗时耗力, 而且影响了信息的传播速度. 因此, 自动检测与判别网络上传播的暴恐音视频成为近年来的一个研究热点.

通常情况下, 对网络暴力元素的判别可以使用视频或音频特征, 也可以两者相结合, 由于音频在处理速度上较快于视频处理速度, 对于实时性要求比较高的场景, 使用音频特征的判别更具优势, 所以基于音频信息的暴恐场景判别研究是极有必要的.

目前暴恐音频场景判别任务主要基于传统的机器学习算法, 采用 SVM 分类器或 KNN 分类器. 2006 年, Theodoros Giannakopoulos 使用 SVM 来检测暴力音频, 提取音频的能量熵、短时平均过零率、短时能量、频谱衰减值等特征, 训练集和测试集各为 10 分钟的视频数据, 最后达到 14.5% 的分类错误率[1]. 文献[2]同样使用 SVM 分类器方法, 在由 15 部电影组成的 MediaEval VSD 数据集上截取枪声、爆炸声等暴力片段, 随机采样 15 部电影中的非暴力部分, 得到 70.2% 的分类准确率. 但由于 SVM 模型在训练数据较多的时候, 需要计算的核矩阵大小也会增大, 将会使训练效率降低, 而较少的训练数据又限制了判别效果. 2008 年, Aggelos Pikrakis 使用 KNN 分类器检测音频中的枪声, 提取了 MFCC、STFT 声谱图、色彩特征、熵、语谱图等特征, 从 30 部电影中截取 5000 个音频片段进行检测, 准确率为 64.55%[3]. 可见采用传统方法进行的暴力音频场景判别效果都不尽人意.

2006 年, Hinton 教授首次提出深度学习的概念, 从此深度学习技术在图像、视频、语音、文本等领域得到了广泛应用. 文献[2]搭建了基于深度神经网络的暴力音频分类系统, 在暴力与非暴力音频各为 30 分钟的训练集上训练, 达到了 77.38% 的分类准确率. 梁嘉欣等人针对传统方法忽略时序信息的问题, 提出了一种基于张量模型的暴力音频检测方法, 最终得到了 89.6% 的准确率[4]. 可见采用了深度学习方法的判别率往往相比于传统机器学习方法有一定提升, 因此本文将卷积神经网络 (Convolutional Neural Network, CNN) 应用于暴恐音频场景的判别中.

由于数据集的限制, 将深度学习用于暴恐音频判别的研究不多. 针对判别网络上传播的一段只有背景信息的音频是否属于暴恐音频的应用背景, 需要含有场景背景信息丰富并且包含暴恐元素的一段持续性音频序列, 然而目前国内外并没有相关领域的公开音频数据集, 于是本文从网上和电影中截取音频片段组成暴恐音频库, 但暴恐音频来源受限并且数量较少, 而 CNN 往往需要希望有足够多的数据训练, 性能才好, 在数据集过小的情况下效果不佳. 于是本文将迁移学习技术引入暴恐音频的判别中.

1 基于迁移学习的暴恐音频判别方法

迁移学习的核心是利用已有的知识, 去解决不同但相关领域的问题[5]. 考虑到本文属于有监督到有监督的类型, 于是采用 fine-tune 的迁移学习方法. Fine-tune 基于一个预训练好的模型, 采用相同的网络结构, 使用不同于预训练好模型的数据, 根据所要完成任务的要求, 调整输出, 在预训练好的模型参数上进行再训练, 是一种解决小数据库训练的方法[6].

图 1 为本文基于迁移学习的暴恐音频判别方法的总体框图, 主要包括提取音频对数梅尔频谱特征、在源音频数据上预训练网络得到预训练模型、在目标音频数据上进行 fine-tune 二次训练. 具体为: 将 TUT 音频数据集作为源音频数据, 提取音频对数梅尔频谱特征后, 预训练网络得到预训练模型, 然后将暴恐音频库作为目标音频数据, 提取对数梅尔频谱特征后在预训练模型上继续训练得到最终的模型, 最后在测试音频上运用最终得到的模型进行暴恐音频判别. 此外, 为提取更多的特征, 在 fine-tune 以后的网络结构中添加辅助网络, 并将辅助网络部分的输出特征与输入特征聚合在一起共同输入分类层, 更有效地利用暴恐音频中的信息.

1.1 对数梅尔频谱特征的提取

音频特征的提取主要有 3 种方式: 时域特征、频域特征及倒谱域特征的提取. 时域特征通常是指短时平均过零率、短时能量、能量熵等, 时域特征具有简单但不够丰富的特点; 频域特征是指傅里叶频谱、滤波器组等. 相比于时域特征, 频域特征具有对外界环境更好的感知特性, 但是频域特征无法得到频率分布随时间变化的状态, 所以本文采用的是音频的倒谱域特

征, 典型代表是对数梅尔频谱特征[7], 将一维的音频信号映射为时间-频域的二维信号[8], 提取过程如图 2 所示.

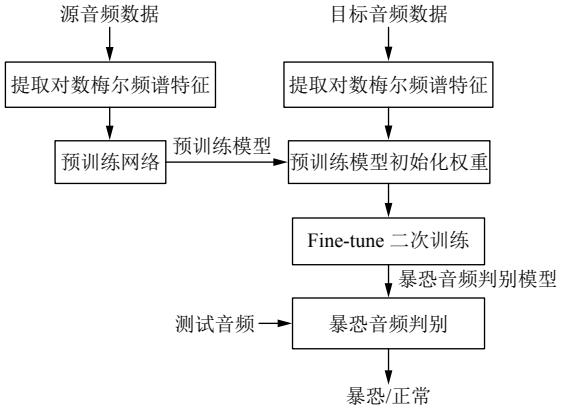


图 1 总体框图

本文产生对数梅尔频谱图的参数为: 音频信号的采样率为 44.1 kHz, 预加重系数为 0.97, 采用汉明窗进行分帧, 快速傅里叶变换 (Fast Fourier Transform, FFT) 窗口长度为 50 ms, 相邻窗之间的距离为 20 ms, 每帧包含 2205 个采样点, 梅尔滤波器的个数为 200, 图 3 展示了含有枪声的暴恐音频 (a) 与正常音频 (b) 的声音波形图与梅尔频谱图.

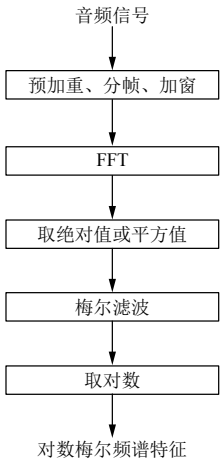
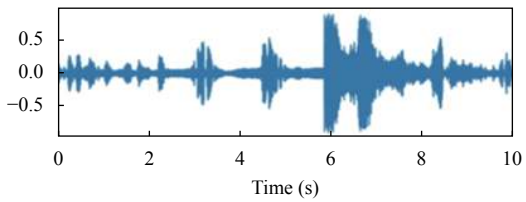
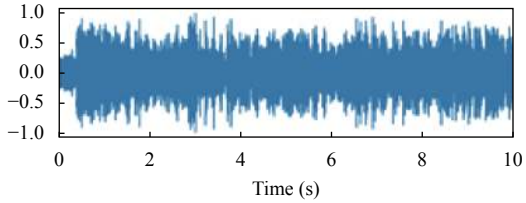


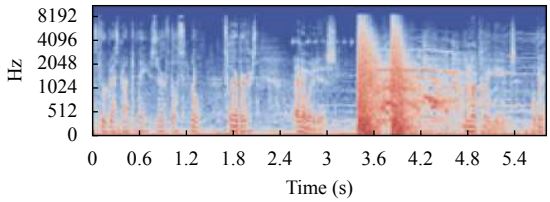
图 2 提取音频对数梅尔频谱特征流程



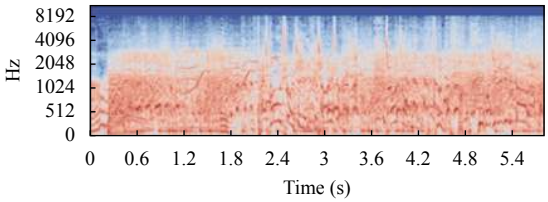
(a) 含有枪声的暴恐音频波形图



(b) 正常音频波形图



(c) 含有枪声的暴恐音频梅尔频谱图



(d) 正常音频的梅尔频谱图

图 3 音频波形图与梅尔频谱图

梅尔频谱图的垂直轴表示频率, 水平轴表示时间, 颜色表示每个时间点各个频率位置处的声音的强度, 图 3(c) 中梅尔频谱图的 3 到 4.2 s 显示的是出现枪声的梅尔频谱图, 与其他未出现枪声的时间段梅尔频谱图有明显差异, 由图 3 可见, 含有暴恐元素音频的频率与强度在整个时间轴上分布不均匀, 而正常音频的梅尔频谱图在整个时间轴上频率与强度分布基本均匀. 提取特征后, 将其转换为分贝标度, 以便于计算.

使用频谱图的好处是把现在的音频分类问题变成了一个图像分类问题, 将每个 wav 文件转换成二维自变量 (时间-频率) 的频谱图, 每个频谱图存储在与其类别相对应的文件夹中. 一个 10 s 长的音频, 采样率为 44.1 kHz, 共有 44.1 kHz×10 s=441 000 个点, 分帧过后, 因为帧移为 20 ms, 即帧移为 882(44.1 kHz×0.02 s=882) 个采样点, 所以维度为 500 列 (441 000/882=500), 行为梅尔滤波器个数. 最终将每个 10 s 长的音频转化为数组形式, 维度为 200 行、500 列.

1.2 预训练网络

在提取音频梅尔频谱特征后, 将每段音频输入卷积神经网络进行预训练, 本文在文献[7]的基础上搭建

预训练网络, 预训练网络结构如图 4.

值得注意的是为了减少神经网络参数与避免过拟合, 采用全局平均池化层 (Global Average Pooling, GAP) 替代全连接层, 搭建的预训练网络结构参数如表 1 所示.

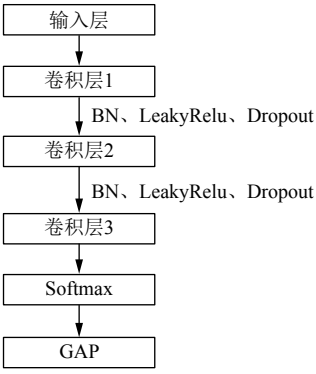


图 4 预训练网络结构

表 1 CNN 模型预训练网络参数表

参数	值
网络结构参数	输入层 (1×200×500) 卷积层 1(100×200×50) 卷积层 2(100×1×1) 卷积层 3(15×1×1)
学习率	0.01
损失函数	交叉熵
Mini Batch	64
迭代次数	600

将上述预训练神经网络在 TUT 数据集上进行实验, 该数据集包含公共汽车站场景类、超市场景类等 15 个场景类, 训练集中每类场景包含 234 个音频, 训练后的模型作为暴恐音频判别的预训练模型.

1.3 基于迁移学习的暴恐音频场景判别方法

由于 TUT 数据集并不包含暴恐音频类, 因此, 需要对预训练模型进行迁移学习, 保留预训练模型权重与网络结构, 调整模型输出, 在自建的暴恐音频库上继续训练, 最终搭建的暴恐音频判别的卷积神经网络结构如图 5.

考虑到 1.2 节得到的预训练模型已经从 TUT 数据集中习得了很多音频低层次特征, 因此只需在预训练模型上做简单权重调整, 所以微调部分迭代次数由 600 改为 300, 学习速率由 0.01 改为 0.001, 模型参数如表 2 所示.

2 改进的 CNN-fine-tune 暴恐音频判别方法

因为在卷积神经网络中, 最终任务的高级特征往往由网络后端习得, 网络前端习得的只是低层次特征[9]. 为提取更多的高级特征, 形成多级特征提取器, 本文以 CNN 模型作为基础网络, 在截断的基础网络的末尾追加了几个特征层, 这部分称为辅助结构. 但如果只是简单地增加深度, 会导致梯度弥散或梯度爆炸, 所以新添辅助结构部分采用了一种类似于残差网络的结构, 如图 6 所示.

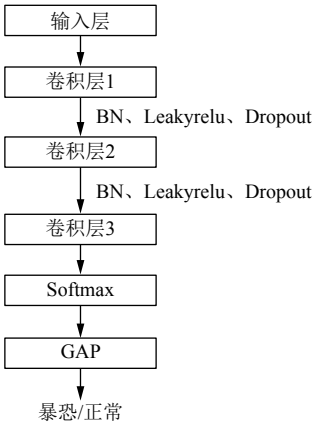


图 5 CNN-fine-tune 网络结构

表 2 CNN-fine-tune 网络参数表

参数	值
网络结构参数	输入层 (1×200×500) 卷积层 1(100×200×50) 卷积层 2(100×1×1) 卷积层 3(2×1×1)
学习率	0.001
损失函数	交叉熵
Mini Batch	64
迭代次数	300

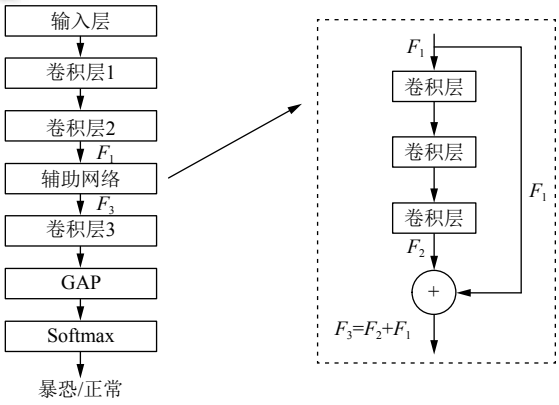


图 6 添加辅助网络结构

辅助网络部分采用 3 个连续的滤波器大小为 1×50、1×1、1×1 的卷积层, 最后一个 1×1 的卷积层为

了改变特征图通道数, 使得经过 3 次非线性激活函数计算, 增强了对于复杂程度和非线性程度的表达能力和泛化能力. 将这一部分得到的特征图与输入特征图聚合在一起, 共同输入分类层:

$$F_2 = \omega_3(\sigma(\omega_2\sigma(\omega_1 F_1))) \tag{1}$$

$$F_2 = concatenate(F_1, F_2) \tag{2}$$

其中, 式 (1) 中F_1是基础网络的输出, 也是辅助网络的输入, F_2是辅助网络的输出, σ为激活函数, ω_i (i = 1,2,3)为辅助结构中 3 个卷积层权重, 采用均匀分布初始化权重. 式 (2) 意为采用 keras 中 *concatenate* 函数, 实现了原始 CNN 网络特征图与辅助网络特征图的数据叠加. 下面说明引入辅助网络的原理[10].

假设o_n是网络第n层的输出特征图, i_n是n层的输入也是第$n-1$层的输出, 每一层输出特征图的计算公式如下:

$$o_n = f_n(i_n, \omega_n, b_n) \tag{3}$$

辅助网络跨越多层, 将输入通过恒等映射转换成输出, 此时每一层的梯度计算公式如下:

$$\frac{\partial o_n}{\partial i_n} = \frac{\partial(i_n + f(i_n\omega_n b_n))}{\partial i_n} = 1 + \frac{\partial f(i_n\omega_n b_n)}{\partial i_n} \tag{4}$$

由式 (4) 可见在网络中加入辅助网络, 可以使得梯度在反向传播时永远大于或等于 1, 这样就不会影响深层网络的训练.

3 实验结果与分析

本文是在 Ubuntu16.04 系统下, 基于 Keras 深度学习框架, 以 theano 作为后端进行网络模型的构建和训练, 实验采用 NVIDIA GTX960 显卡进行加速.

预训练网络部分使用 TUT 数据集, 迁移学习网络部分训练与测试数据集组成如下: 从 Youtube 中下载了网友录制的一些恐怖袭击现场音频, 同时也选取了少部分电影中的暴恐镜头音频, 根据枪声、尖叫声、爆炸声、警报声、打斗声等截取音频. 正常音频包括综艺节目片段、电影片段与生活场景音频, 包含了笑声、说话声、鼓掌声、音乐声等. 建立的数据集共 699 个音频片段, 由于音频段时长各异, 制作数据集时统一将尺寸设定为每个音频 10 s, 其中正常音频片段 348 个, 250 个正常音频用于训练, 98 个正常音频用于测试, 暴恐音频片段 351 个, 250 个暴恐音频用于训练, 101 个暴恐音频用于测试. 音频库分布如表 3.

表 3 音频分布表

数据集	正常音频		暴恐音频	
	个数	时长 (min)	个数	时长 (min)
训练集	250	41.67	250	41.67
测试集	98	16.33	101	16.83

本文研究的对象是一个二分类任务, 各种类别样本数量相当, 不需要考虑样本类别不平衡问题, 性能指标采用准确率 (Accuracy, *Acc*). *Acc* 的计算过程如公式所示:

$$Acc = \frac{N_t}{N_{\text{all}}} \tag{5}$$

其中, N_t表示每类预测正确的样本数量, N_{all}表示每类总样本数量.

利用最终得到的暴恐音频判别模型在测试集的 199 个音频片段上进行测试, 得到未使用迁移学习与使用迁移学习, 以及未改进 CNN 与改进 CNN 后得到的判别效果分别如表 4 所示, 同时使用传统 SVM 分类器进行比较.

表 4 实验结果

实验	方法	*Acc*(%)		
		暴恐音频	正常音频	平均判别率
实验一	SVM 分类器	87.13	33.67	60.4
实验二	未使用 fine-tune	90.10	93.88	91.99
实验三	使用 fine-tune	97.03	94.90	95.96
实验四	改进的迁移学习方法	98.02	95.92	96.97

由实验一结果与实验二结果对比可得, 传统机器学习方法对于暴恐音频的判别不如深度学习方法. 实验二和实验三对比可得, 使用 fine-tune 的迁移学习方法比未使用迁移学习的方法提升了 6.93% 的暴恐音频判别率和 1.02% 的正常音频判别率, 平均判别率提升了 3.97%. 同时, 实验四表明叠加辅助网络结构后对于暴恐音频和正常音频的判别率都有所提高, 平均判别率相比于未添加辅助网络的提高了 1.01%, 可见叠加的辅助网络有助于得到更加可靠的特征提取效果.

4 结论

本文在判别网络上传播的一段音频是否属于暴恐音频的应用背景下, 首先在公开的 TUT 音频数据集上进行预训练得到预训练模型, 然后利用 fine-tune 的迁移学习方法将预训练模型与网络结构引入暴恐音频的判别中, 在训练数据较少的情况下, 也得到了不错的判

别率, 为提取更多的特征, 在 fine-tune 以后的网络添加了一种类似于残差网络的结构, 进一步提高了音频判别率.

参考文献

1 Giannakopoulos T, Kosmopoulos D, Aristidou A, *et al*. Violence content classification using audio features. Proceedings of the 4th Helenic Conference on Artificial Intelligence. Crete, Greece. 2006. 502–507.

2 冯佳军. 暴力音频场景分类技术研究与系统实现[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2016.

3 Pikrakis A, Giannakopoulos T, Theodoridis S. Gunshot detection in audio streams from movies by means of dynamic programming and Bayesian networks. Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, NV, USA. 2008. 21–24.

4 梁家欣, 李海峰, 马琳. 基于张量模型的暴力音频检测研究. 智能计算机与应用, 2016, 6(1): 108–111. [doi: 10.3969/j.issn.2095-2163.2016.01.030]

5 王红斌, 沈强, 线岩团. 融合迁移学习的中文命名实体识别. 小型微型计算机系统, 2017, 38(2): 346–351.

6 孙超, 吕俊伟, 刘峰, 等. 基于迁移学习的红外图像超分辨率方法研究. 激光与红外, 2017, 47(12): 1559–1564. [doi: 10.3969/j.issn.1001-5078.2017.12.019]

7 Piczak K J. The details that matter: Frequency resolution of spectrograms in acoustic scene classification. Proceedings of the Detection and Classification of Acoustic Scenes and Events. Munich, Germany. 2017. 1–5.

8 周成豪. 基于概率潜在语义分析的音频场景识别方法[硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2013.

9 关胤. 基于残差网络迁移学习的花卉识别系统. 计算机工程与应用, 2019, 55(1): 174–179.

10 张振焕, 周彩兰, 梁媛. 基于残差的优化卷积神经网络服装分类算法. 计算机工程与科学, 2018, 40(2): 354–360. [doi: 10.3969/j.issn.1007-130X.2018.02.023]