

文章编号:1007-5321(2020)05-0112-06

DOI:10.13190/j.jbupt.2020-075

一种基于 EEMD 的异常声音识别方法

韦娟¹, 顾兴权¹, 宁方立²

(1. 西安电子科技大学 通信工程学院, 西安 710071; 2. 西北工业大学 机电学院, 西安 710072)

摘要: 为了优化组合特征在异常声音识别中的效率,提出一种用集合经验模态分解(EEMD)对异常声音帧信号进行有效性检测和提取多层特征的算法. 首先对异常声音帧信号进行集合经验模态分解,得到固有模态函数;然后根据给定的固有模态函数层数阈值,对该帧信号进行有效性检测;再对有效帧信号的每一层固有模态函数提取梅尔频率倒谱系数、翻转梅尔频率倒谱系数、线性预测倒谱系数、短时能量和能量比,并将它们归一化后拼接成多层特征. 根据提取的特征,用深度卷积神经网络实现异常声音识别分类. 仿真结果表明,提出的新方法在4类异常声音识别中的识别率可以达到98.65%.

关键词: 异常声音识别; 集合经验模态分解; 多层特征; 深度卷积神经网络

中图分类号: TP391

文献标志码: A

An Abnormal Sound Recognition Method Based on EEMD

WEI Juan¹, GU Xing-quan¹, NING Fang-li²

(1. School of Communication Engineering, Xidian University, Xi'an 710071, China;

2. School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: In order to optimize the efficiency of combined features in abnormal sound recognition, an algorithm for detecting the effectiveness of abnormal sound frame signals and extracting multi-layer features using ensemble empirical mode decomposition (EEMD) is proposed. Firstly, an ensemble empirical mode decomposition is performed on the abnormal sound frame signal to obtain the intrinsic model function, and then the validity of the frame signal is tested according to the given layer threshold of the intrinsic modal function. Finally, the Mel frequency cepstral coefficients, the inverted Mel frequency cepstral coefficients, the linear prediction cepstral coefficients, the short-time energy and energy ratio are extracted for each layer of the intrinsic modal function of the effective frame signal, and then all of them are normalized and spliced into multi-layer feature. According to the extracted features, the deep convolutional neural network is used to realize the classification and recognition of abnormal sound. Simulations show that the proposed new method can achieve a recognition rate of 98.65% in four types of abnormal sound recognition.

Key words: abnormal sound recognition; ensemble empirical mode decomposition; multi-layer feature; deep convolutional neural network

随着社会的进步,人们对自身安全的要求不断提高,尤其是在公共场所. 公共场所发生的异常事件通常都伴随着枪声、爆炸声和玻璃破碎声等异常

声音. 为了有效提高异常事件发生后的快速响应能力,对异常声音进行精准识别具有非常重要的现实意义.

收稿日期: 2020-06-25

基金项目: 国家自然科学基金项目(51675425); 陕西省重点研发计划项目(2018GY-181, 2020ZDLGY06-09)

作者简介: 韦娟(1973—), 女, 副教授, E-mail: weijuan@xidian.edu.cn.

异常声音是一种典型的非平稳信号,陈等^[1]用集合经验模态分解(EEMD, ensemble empirical mode decomposition)提取异常声音的固有模态函数(IMF, intrinsic mode function);然后对每一层IMF提取时域和频域特征;最后取它们的平均值作为识别特征,但取平均在一定程度上会破坏特征的固有规律. Xu等^[2]以梅尔频率倒谱系数(MFCC, Mel frequency cepstral coefficients)和短时能量为特征,用高斯混合模型对3种环境的异常声音进行识别,结果表明,组合特征的识别效果优于单一特征的识别效果,但识别率会受样本数量的影响. Pedroza等^[3]以MFCC和翻转梅尔频率倒谱系数(IMFCC, inverted Mel frequency cepstral coefficients)为特征,对鸟类声音进行识别,发现IMFCC在鸟类声音识别中表现的效果更好,但识别率与鸟的类别有关,泛化性不强. 李等^[4]提出一个玻璃破碎声音的实时识别系统,用小波包分析提取玻璃破碎声音的时域和频域特征,然后用隐马尔可夫模型进行识别,但识别时需要将玻璃破碎声音的频段进行划分. Yan等^[5]提出一种在低信噪比条件下,融合时域和频域特征,对声音事件进行识别的方法,虽然融合后的高维特征能够提高识别率,但识别率依旧不够理想. 韦娟等^[6]使用EEMD提取异常声音的MFCC、短时能量和短时能量比,然后组合起来验证所提方法的优劣性. 虽然能够提高异常声音的识别率,但是需要对分类器进行复杂的参数寻优.

考虑到组合时域和频域特征的优势,以优化组合特征为切入点,提出一种基于EEMD的多层特征提取方法. 首先对异常声音帧信号做EEMD,计算各阶IMF与该帧信号的相关系数;然后根据相关系数阈值筛选出更有效的IMF;再对筛选到的每一层IMF提取MFCC、IMFCC、线性预测倒谱系数(LPCC, linear prediction cepstral coefficients)、IMF短时能量和IMF能量比,并且直接归一化后拼接成一个多层特征;最后用多层特征训练深度卷积神经网络,对异常声音进行识别分类.

1 有效异常声音检测

EEMD以时间复杂度为代价,能够优化经验模态分解(EMD, empirical mode decomposition)中存在的模态混叠效应,常取代EMD对非平稳信号进行分解. 由于EMD算法具有正交性^[7],可以通过计算各阶IMF分量同原始信号间的相关系数 ρ_i ,并设定选

择阈值的方式,提取对原始信号较为敏感的IMF分量^[8]. 为了避免用于特征提取的异常声音中含有静音段,或者夹杂其他冗余信号,导致提取的特征效率低的问题,在文献[8]的基础上,根据异常声音信号 $x(t)$ 自身特性设定相应的IMF层数阈值,对 $x(t)$ 进行有效性检测. 各层IMF的相关系数为

$$\rho_i = \frac{\sum_{t=0}^N \sum_{k=1}^l y_k(t)x(t)}{\sqrt{\sum_{t=0}^N \sum_{k=1}^l y_k^2(t)x^2(t)}}, i=1,2,\dots,l \quad (1)$$

其中: $y_k(t)$ 为第 k 阶IMF的时域信号, N 为 $x(t)$ 的长度, l 为IMF的阶数.

有效异常声音检测包括以下5个步骤.

1) 异常声音信号 $x(t)$ 分帧加窗,窗函数为汉明窗,帧长和帧移分别为512和256个采样点.

2) 确定IMF层数阈值 θ ,用时间复杂度较小的EMD对任意1 000帧异常声音信号进行分解,并统计分解所得IMF的层数,然后依据多数原则选择数量最多的层数作为阈值 θ .

3) 根据 θ 值,对每一帧异常声音信号 $x_i(t)$ 做步骤4)和步骤5)的操作.

4) 对 $x_i(t)$ 做EEMD分解,并统计分解所得IMF的层数 l .

5) 若 $l \geq \theta$,则记 $x_i(t)$ 为有效的异常声音帧信号,然后按照式(1)计算各层IMF的相关系数 ρ_i ,并按从大到小的顺序排列,再选择 ρ_i 最大的前 θ 层IMF用于特征提取,而其余的IMF被认为与该帧信号相关性不大,不用于特征提取;若 $l < \theta$,则认为 $x_i(t)$ 为静音帧或者冗余帧,并丢弃该帧.

2 多层特征提取

2.1 特征选择

MFCC是在Mel标度频率域提取出来的频域特征^[9],常用于异常声音识别. 但由于Mel标度在高频区域分布稀疏,MFCC对高频信息的表征不够完全. Chakroborty等^[10]提出了与MFCC互补的频域特征IMFCC,以补充提取MFCC时丢失的部分高频信息. LPCC也是异常声音识别中一种常用的频域特征^[11],具有可靠性高和鲁棒性强的特点,通常采用全极点模型^[12]对线性预测系数进行递推得到. 异常声音具有短时爆发性强的特点,所以结合时域特征的优势以及IMF的频率差异性,提取反映信号时域能量变化的IMF层能量 E_{IMF} 和反映不同频率尺

度能量分布的 IMF 层能量比 σ , 有

$$E_{\text{IMF}} = \sum_{t=1}^N y^2(t) \quad (2)$$

$$\sigma = \frac{E_{\text{IMF}}}{E_x} \quad (3)$$

其中 E_x 为对应帧信号 $x_i(t)$ 的短时能量。

2.2 特征提取

组合时域和频域特征是异常声音识别中特征选择的一个趋势,且不同特征在同一异常声音识别中表现出来的效果不同^[13]。于是,基于文献[1]的特征提取思路,对有效异常声音帧信号 $x_i(t)$ 的每一层 IMF 提取时域和频域特征,但为了充分保留每一层 IMF 的固有特性,直接将提取的特征值归一化后拼接成一个高维的多层特征。由于目前还没有比较系统的异常声音音频库,所以从 China Webmaster 网站下载了枪声、爆炸声、玻璃破碎声和脚步声作为待分类异常声音,采样频率为 22.05 kHz,单声道,长度为

1~8 s不等^[6]。

多层特征的提取包括 5 个步骤。

1) 确定 IMF 层数阈值 θ , 枪声、爆炸声、玻璃破碎声和脚步声的 θ 值分别为 7、6、8 和 7。

2) 根据 θ 值对 $x_i(t)$ 进行有效性检测,获得 θ 层 IMF,然后按图 1 所示流程提取 θ 层特征,即对每一层 IMF 做步骤 3) 和步骤 4) 的操作。

3) 提取 12 维 MFCC、IMFCC 和 LPCC,并将特征值归一化到 $[-1, 1]$;按照式(2)和式(3)计算 IMF 层能量 E_{IMF} 和 IMF 层能量比 σ ,并将 E_{IMF} 归一化到 $[-1, 1]$ 。

4) 依次拼接特征 MFCC、IMFCC、LPCC、 E_{IMF} 和 σ ,形成大小为 (1×38) 的特征向量。

5) 按照有效性检测时 IMF 的排列次序,依次拼接每一层 IMF 的特征向量,形成大小为 $(\theta \times 38)$ 的特征向量,然后将该特征向量直接可视化成一个图片,即一个 θ 层特征。

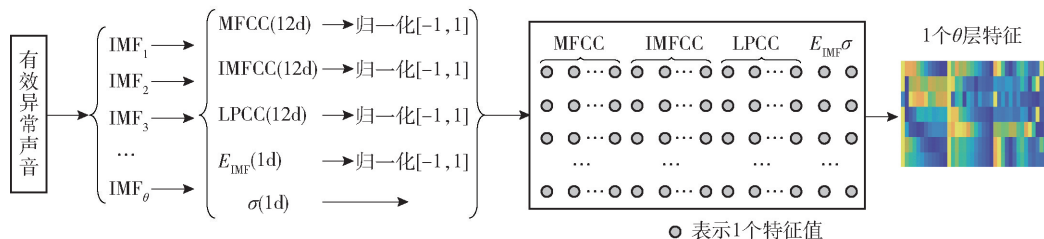


图1 一个 θ 层特征提取流程

3 分类模型

一般机器学习方法使用的特征是单维等长的,故多层特征不再适用,所以可选择能对图片进行分类识别的深度卷积神经网络。Inception-v3^[14]模型应用卷积核分解思想,将 3×3 卷积核用 3×1 卷积

核和 1×3 卷积核替换,能够降低计算量和参数量,是现有深度学习模型中表现较为突出的一个。所以,可采用如图 2 所示的深度卷积神经网络进行异常声音分类识别,图中表达式 $(m \times n)/s$ 表示该层的卷积核大小为 $(m \times n)$,步长为 s ,而 $x \times y \times z$ 表示该层的输入尺寸。

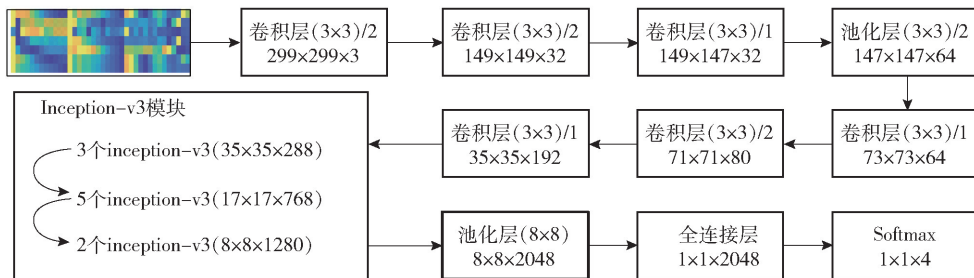


图2 分类模型

4 仿真实验与结果分析

多层特征提取部分在 Matlab 2018b 平台上完成,识别分类部分在基于 python3.5 的 Tensor-

Flow1.1.0 环境下完成。硬件配置为:2.8 GHz Intel i5 处理器,8 GB 1 600 MHz DDR3L 内存,Nvidia GeForce GTX 860 显卡。每一类异常声音提取 2 500 个多层特征样本,其中 2 000 个作为训练样本,500 个

作为测试样本,故训练集样本总数为 8 000 个,测试集样本总数为2 000 个. 实验中,学习率为 0. 01,批大小为 10,训练步数为 1 万步. 模型衡量指标为测试集的识别准确度,即测试集样本的识别结果与样本的标签相吻合的样本占全部样本个数的百分比.

4. 1 不同特征提取条件对准确度的影响

为了说明有效性检测和多层特征的层数 θ 对异常声音识别分类准确度的影响,提取以下 7 组特征:不进行有效性检测,直接对异常声音帧信号提取的特征 F_{nc} ;当 4 类异常声音的 θ 值分别等于 7、6、8 和 7 时提取的特征 F_d ;当四类异常声音的 θ 值均为 6、5、4、3 和 2 时提取的特征 F_6, F_5, F_4, F_3 和 F_2 . 表 1 所示为这 7 组特征训练的模型在测试集上的准确度随训练步数 n 的变化情况以及模型训练 1 万步时所需的总时间和单个样本的平均预测时间.

表 1 不同特征提取条件下模型的准确度和相关时间的变化情况

特征	训练步数/万步					总时间/h	单个样本平均预测时间/ms
	0. 2	0. 4	0. 6	0. 8	1. 0		
F_{nc}	26. 05	39. 40	53. 85	25. 00	25. 36	4. 50	28. 90
F_d	55. 55	48. 70	72. 10	83. 25	98. 65	4. 92	28. 63
F_6	25. 00	35. 45	54. 35	79. 15	87. 50	5. 67	28. 79
F_5	24. 55	27. 80	79. 30	76. 10	81. 55	6. 32	28. 66
F_4	25. 00	25. 65	68. 95	64. 50	77. 45	6. 83	28. 97
F_3	25. 00	37. 75	63. 40	75. 15	73. 20	5. 67	28. 91
F_2	29. 55	45. 95	32. 85	55. 65	74. 90	5. 28	28. 58

从表 1 可知,有效性检测能够提高模型的训练准确度,用特征 F_{nc} 训练的模型其最大准确度仅为 53. 85%,且随着训练步数增加,发生严重的过拟合,但使用经过有效性检测的特征训练的模型的最小准确度便可以达到 74. 9%. 此外,随着训练步数的增加,准确度总体呈上升趋势,使用特征 F_d 仅训练 1 万步,准确度可以达到 98. 65%,说明多层特征是一种有效的特征,且用多层特征训练模型时,模型的收敛速度比较快. 但是,多层特征的层数会严重影响准确度的大小,且 θ 的取值越大,准确度的值就越大. 使用特征 F_d 和 F_6 训练模型的准确度相差 11. 15%,说明根据异常声音自身特性选取 θ 的差别越大,即各类异常声音的频率分布差异越大,提取的多层特征的区分效果就会越好. 使用特征 F_6, F_5, F_4, F_3 和 F_2 训练模型准确度的最大值分别为 87. 50%、81. 55%、77. 45%、

75. 15% 和 74. 90%,说明 θ 的取值越大,多层特征的表现效果会越好. 由表 1 还可知,不同特征训练模型所需时间不同,但使用训练好的模型识别一个样本的时间基本相同. 使用特征 F_d 和 F_4 训练模型所需时间的差值为1. 91 h,但单一样本的预测时间差值仅为 0. 34 ms.

表 2 所示为多层特征在测试集上获得最大准确度时模型在每一类信号上的准确度.

表 2 不同模型下每一类信号的准确度

特征	枪声	爆炸声	破碎声	脚步声
F_d	98. 97	100. 00	100. 00	96. 30
F_6	78. 94	82. 45	92. 13	97. 81
F_5	60. 52	90. 78	82. 27	99. 60
F_4	53. 92	82. 70	92. 11	98. 79
F_3	53. 38	79. 32	91. 62	99. 78
F_2	52. 03	69. 76	87. 70	94. 92

分析表 2 可知,多层特征对脚步声的准确度影响较小,对枪声、爆炸声和玻璃破碎声的准确度影响较大. 使用表 2 中的 6 组特征训练的模型识别枪声、爆炸声和玻璃破碎声,准确度的最大值和最小值的差值分别是 46. 94%、30. 24% 和 17. 73%,但用相同的模型识别脚步声,差值仅为 4. 86%. 此外,多层特征的层数会影响模型的均衡性,使用特征 F_2 和 F_6 训练的模型在 4 类异常声音上的准确度的差值分别为 42. 89% 和 18. 87%,但使用特征 F_d 训练模型的差值仅为 3. 7%.

4. 2 信噪比不同时提取的特征

为了验证多层特征在异常声音识别中对噪声的鲁棒性,在 θ 等于 6 时,直接向异常声音中加入高斯白噪声,形成信噪比 (SNR, signal-to-noise ratio) 等于 0、5、15 和 30 dB 的带噪信号,并提取相应的特征 $F_{0\text{ dB}}, F_{5\text{ dB}}, F_{15\text{ dB}}$ 和 $F_{30\text{ dB}}$. 表 3 所示为这 4 组特征训练模型下的准确度随训练步数 n 的变化情况.

表 3 不同 SNR 条件下模型的准确度变化情况

特征	2 000 步	4 000 步	6 000 步	8 000 步	1 万步
$F_{0\text{ dB}}$	25. 05	25. 10	48. 65	47. 30	69. 55
$F_{5\text{ dB}}$	25. 00	24. 49	46. 15	58. 80	71. 75
$F_{15\text{ dB}}$	28. 25	25. 00	45. 10	66. 70	78. 65
$F_{30\text{ dB}}$	37. 80	52. 60	50. 25	73. 40	83. 15

分析表 3 可知,在有噪声的情况下,多层特征依旧是一种有效、且收敛速度比较快的特征,训练 1 万

步,特征 $F_{0\text{ dB}}$, $F_{5\text{ dB}}$, $F_{15\text{ dB}}$ 和 $F_{30\text{ dB}}$ 所能获得的最大准确度分别为 69.55%、71.75%、78.65% 和 83.15%。随着 SNR 的增加,准确度的最大值也不断增加,SNR 等于 0 时的最大准确度仅比 SNR 等于 30 dB 时的最大准确度低 13.6%,而 SNR 等于 30 dB 时的准确度仅比没有加噪时的准确度低 4.35%,说明多层特征对噪声具有很好的鲁棒性,且多层特征在异常声音识别中有比较好的稳定性。

5 结束语

为了优化组合特征在异常声音识别中的效率,基于 EEMD 提取异常声音的时域和频域特征,并将提取的低维特征加工成表现更直接的高维多层特征。通过提取不同层数的多层特征和不同信噪比下的多层特征,训练深度卷积神经网络实现异常声音识别分类,得出以下结论:

1) 多层特征对层数变化比较敏感,当待分类的各异常声音的频率分布范围差别比较大时,多层特征的层数变化会比较明显,相应的识别率会更高,但是如何更好地确定多层特征的层数还需要进一步研究;

2) 多层特征的层数越高,所包含的特征值越多,识别效果会越好,且用多层特征训练模型的收敛速度比较快;

3) 多层特征是一种相对比较稳定的组合特征,且对噪声的鲁棒性比较好。但笔者只将其运用到 4 类异常声音识别中,后期还需拓展到更多类异常声音识别中。

参考文献:

- [1] 陈志全,杨骏,乔树山. 基于 EEMD 的异常声音特征提取[J]. 计算机与数字工程, 2016, 44(10): 1875-1879.
Chen Zhiquan, Yang Jun, Qiao Shushan. Abnormal sound feature extraction based on EEMD[J]. Computer and Digital Engineering, 2016, 44(10): 1875-1879.
- [2] Xu Jining, Yao Xiaoxin. Abnormal sound recognition with audio feature combination and modified GMM[C]// Proceedings of the 32nd Chinese Control Conference. Xi'an: IEEE Press, 2013: 4582-4585.
- [3] Pedroza Ramirez A D, De La Rosa Vargas J I, Valdez R R, et al. A comparative between Mel frequency cepstral coefficients and inverse Mel frequency cepstral coefficients features for an automatic bird species recognition system[C]//2018 IEEE Latin American Conference on Computational Intelligence. Guadalajara: IEEE Press, 2018.

- [4] 李颀,白雨尼,王丹聪. 基于小波包分析的玻璃破碎声音识别系统设计[J]. 计算机测量与控制, 2018, 26(1): 168-172.
Li Qi, Bai Yuni, Wang Dancong. Design of class breaking sound recognition system based on wavelet packet transform[J]. Computer Measurement and Control, 2018, 26(1): 168-172.
- [5] Yan Guolin, Wang Mei, Liu X, et al. Sound event recognition based in feature combination with low SNR[C]//2019 International Conference on Artificial Intelligence and Advanced Manufacturing. Dublin: IEEE Press, 2019: 109-114.
- [6] 韦娟,张芃楠,岳凤丽,等. 基于 PSO-PF 算法的 SVM 识别方法及其在异常声音中的应用[J]. 北京邮电大学学报, 2019, 42(3): 58-63.
Wei Juan, Zhang Pengnan, Yue Fengli, et al. Recognition and application of abnormal sound via SVM based on PSO-PF[J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(3): 58-63.
- [7] Zhang Shangyue, Liu Yuanyuan, Yang Gongliu. EMD interval thresholding denoising based on correlation coefficient to select relevant modes[C]//2015 34th Chinese Control Conference. Hangzhou: IEEE Press, 2015: 4801-4806.
- [8] 杨恭勇,周小龙,李家飞,等. 局部 Hilbert 边际能量谱在滚动轴承故障诊断中的应用[J]. 东北电力大学学报, 2017, 37(2): 77-81.
Yang Gongyong, Zhou Xiaolong, Li Jiafei, et al. A study of rolling bearing fault diagnosis based on local Hilbert marginal energy spectrum[J]. Journal of Northeast Electric Power University, 2017, 37(2): 77-81.
- [9] Winursito A, Hidayat R, Beji A, et al. Improvement of MFCC feature extraction accuracy using PCA in indonesian speech recognition[C]//2018 International Conference on Information and Communications Technology. Yogyakarta: IEEE Press, 2018: 379-383.
- [10] Chakroborty S, Roy A, Majumdar S, et al. Capturing complementary information via reversed filter bank and parallel implementation with MFCC for improved text-independent speaker identification[C]//2007 International Conference on Computing: Theory and Applications. Kolkata: IEEE Press, 2007.

- [11] Gupta H, Gupta D. LPC and LPCC method of feature extraction in speech recognition system [C] // 2016 6th International Conference Cloud System and Big Data Engineering. Noida; IEEE Press, 2016: 498-502.
- [12] Eltiraifi O, Elbasheer E, Nawari M. A comparative study of MFCC and LPCC features for speech activity detection using deep belief network [C] // 2018 International Conference on Computer Control, Electrical and Electronics Engineering. Khartoum; IEEE Press, 2018.
- [13] Liu Gang, He Wei, Jin Bicheng. Feature fusion of speech emotion recognition based on deep learning [C] // 2018 International Conference on Network Infrastructure and Digital Content. Guiyang; IEEE Press, 2018: 193-197.
- [14] Nikhitha M, Roopa Sir S, Uma Maheswari B, et al. Fruit recognition and grade of disease detection using inception V3 model [C] // 2019 3rd International Conference on Electronics, Communication and Aerospace Technology. Coimbatore; IEEE Press, 2019: 1040-1043.

(上接第 104 页)

- [7] Ashish V, Noam S, Niki P, et al. Attention is all you need [C] // Proc of the 30th Advances in Neural Information Processing Systems (NIPS). Cambridge; MIT Press, 2017: 6000-6010.
- [8] Luo Ruixuan, Xu Jingjing, Zhang Yi, et al. PKUSEG: a toolkit for multi-domain Chinese word segmentation [EB/OL]. 2019(2019-06-27) [2020-07-18]. <https://arxiv.org/abs/1906.11455>.
- [9] Tomas M, Kai C, Greg C, et al. Efficient estimation of word representations in vector space [C] // Proc of the 1rd Int Conf on Learning Representations (ICLR). Scottsdale; ICLR, 2013: 1-12.
- [10] Reuven Y. Rubinstein. Optimization of computer simulation models with rare events [J]. European Journal of Operational Research, 1997, 99(1): 89-112.
- [11] GitHub. DFF-Dataset [EB/OL]. 2020 (2020-06-18) [2020-10-04]. <https://github.com/liuyichenaal/DFF-Dataset>.
- [12] Adam P, Sam G, Francisco M, et al. pyTorch: an imperative style, high-performance deep learning library [C] // Proc of the 32th Advances in Neural Information Processing Systems (NeurIPS 2019). Vancouver; MIT Press, 2019: 8024-8035.
- [13] Guo Long, Zhang Dongxiang, Wang Lei, et al. CRAN: a hybrid CNN-RNN attention-based model for text classification [C] // Proc of the 37th International Conference on Conceptual Modeling (ER2018). Berlin; Springer-Verlag, 2018: 571-585.
- [14] Jiang Wei, Jin Zhong. Integrating bidirectional LSTM with inception for text classification [C] // Proc of the 4th IAPR Asian Conference on Pattern Recognition (ACPR 2017). Piscataway; IEEE, 2017: 870-875.