使用基于CNN的特征自行检测异常声音监督学习

技术报告

森田一树⋆, 矢野智彦⋆, 凯 Q. 陈⋆

SECOM CO.,LTD. 智能系统实验室
{morita-ka,tomo-yano,ku-chan}@secom.co.jp

抽象的

我们在本报告中为 DCASE2021 task2 的异常声音检测任务提出了一种检测方法。这就是机器状态监测的异常声音检测任务，要求仅从正常声音数据中检测出未知的异常声音。我们使用机器的正常声音及其截面索引以自监督学习的方式训练卷积神经网络（CNN）。然后，我们使用从 CNN 中提取的特征向量来检测异常声音。因此，对于开发数据集，我们显示曲线下面积 (AUC) 的检测性能为 78.05%，部分 AUC (pAUC) 的检测性能为 68.09%。

索引词— 异常声音检测，卷积神经网络

一、介绍

在DCASE2021task2“Domain Shifted Condition下机器状态监测的无监督异常声音检测”[1]中，要求检测机器的异常声音。由于我们只能获取机器的正常声音，因此异常声音检测是一个无监督的问题。在 DCASE2021 任务中，新增了训练数据和测试数据的声学特性不同的条件（即 domain shift）。

在 DCASE2020 task2 中，我们只使用了常规的检测方法，我们发现提取更有效的特征很重要。因此，在 DCASE2021 task2 中，我们基于卷积神经网络（CNN）从机器的声音中提取特征。此外，我们使用与去年相同的常规异常检测方法。

本文组织如下。在第 2 章中，我们描述了我们的异常声音检测方法。在第 3 章中，我们展示了评估实验和结果。在第 4 章中，我们总结了本报告。在第 5 章中，我们描述了我们提交的模型。

2. 异常声音检测方法

2.1. 音频处理

我们将所有音频剪辑转换为频谱图。STFT 的帧大小为 128 毫秒，跳大小为 32 毫秒。我们通过实验设置这些参数。我们使用频谱图作为 CNN 的输入。

⋆平等贡献。

2.2. 使用 CNN 的特征提取器

通过使用频谱图和截面索引，我们训练了一个 CNN，例如 MobileNetV2(MNv2)[2] 和 MobileFaceNet(MFN)[3]。此外，我们使用 Additive Angular Margin Loss [4] 作为损失函数。1024维的频谱图 × 以32帧为处理单位，单位在音频剪辑中移位16帧。每个模型结构如表 1 和表 2 所示。因此，我们获得了每单位 128 维的向量。

表 1：MobileNetV2 架构

输入	操作员	吨	C	n	秒
1024 ×32 ×1	conv2d 3 ×3	-	32	1	2
512 ×16 ×32	瓶颈	1	16	1	1
512 ×16 ×16	瓶颈	6	24	2	2
256 ×8 ×24	瓶颈	6	32	3	2
128 ×4 ×32	瓶颈	6	64	4	2
64 ×2 ×64	瓶颈	6	96	3	1
64 ×2 ×96	瓶颈	6	160	3	2
32 ×1 ×160	瓶颈	6	320	1	1
32 ×1 ×320	conv2d 1 ×1	-	1280	1	1
32 ×1 ×1280	大街 16 号池 ×1	-	-	1	——
1 ×1 ×1280	conv2d 1 ×1	-	128	-	

表 2：MobileFaceNet 架构

输入	操作员	吨	C	n	秒
1024 ×32 ×1	conv2d 3 ×3	-	64	1	2
512 ×16 ×64	深度conv2d 3 ×3	-	64	1	1
512 ×16 ×64	瓶颈	2	64	5	2
256 ×8 ×64	瓶颈	4	128	1	2
128 ×4 ×128	瓶颈	2	128	6	2
64 ×2 ×128	瓶颈	4	128	1	2
32 ×1 ×128	瓶颈	2	128	2	1
32 ×1 ×128	conv2d 1 ×1	-	512	1	1
32 ×1 ×512	线性 GDConv16 ×1	-	512	1	1
1 ×1 ×512	线性 conv2d 1 ×1	-	128	1	1

2.3. 异常检测器

我们对源域应用局部异常值因子（LOF），对目标域应用 k-最近邻（k-NN）。我们使用均值或标准差合并音频剪辑中的嵌入向量。

局部异常因子（LOF）[5]

该方法基于局部密度，即相邻特征值的密度。当特征异常时，异常的局部密度与相邻特征之间的差异很大。在本报告中，我们使用 LOF 的输出作为异常分数。我们将邻居的数量设置为 4。

k-最近邻（k-NN）[6]

该方法基于k-相邻特征的距离。在k-NN中，到所选邻域的距离越大，偏离正常的越多。在本报告中，我们使用余弦距离的平均值作为异常分数，并将邻居数设置为 1。

3. 评估实验

3.1. 实验条件

10 秒长的音频（单声道，16 kHz）是从机械声源中采样的。机器有七种类型（Machine Type）；ToyCar、ToyTrain[7]、风扇、变速箱、泵、滑块和阀门[8]。对于每种机器类型，开发数据集中有 3 个部分，附加数据集中有 3 个部分。我们使用机器类型中的 6 个部分数据集训练了一个 CNN，并使用每个部分的嵌入向量训练了一个异常检测器。我们使用 librosa[9] 和 scikit-learn[10] 来实现。当我们在源域中评估声音片段时，我们只使用源域中的训练数据，而在目标域中时，我们使用源域和目标域中的训练数据。

在实验中，我们比较了以下内容：

- 特征提取器模型：MobileNetV2、MobileFaceNet
- 异常检测器模型：LOF、k-NN
- 特征合并方法：均值、标准差(std)

3.2. 结果

结果如表3、表4、表5和表6所示。源域的结果如表3和表4所示，目标域的结果如表5和表6所示。每个值是 AUC 或 pAUC 整体部分的调和平均值。

4。结论

在本文中，我们使用机器的正常声音及其截面索引以自监督学习的方式训练CNN。然后，我们使用从 CNN 中提取的特征向量来检测异常声音。开发数据集显示了 AUC 的 78.05% 和 pAUC 的 68.09% 的性能。

5. 提交

在本报告中，我们提交了三个异常声音检测系统。表 7 显示了我们使用的条件。

6. 参考资料

[1] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi、R. Tanabe、H. Purohit 和 T. Endo，“DCASE 2021 挑战任务 2 的描述和讨论：域转移条件下机器状态监测的无监督异常声音检测，” *在 arXiv 电子版中：2106.04492, 1 – 5*, 2021.

[2] M. Sandler、A. Howard、M. Zhu、A. Zhmoginov 和 L.-C. Chen，“MobileNetV2：倒置残差和线性瓶颈”，第 4510-4520 页，2018 年。

[3] S. Chen、Y. Liu、X. Gao 和 Z. Han，“MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices”，第 428-438 页，2018 年。

[4] J. Deng、J. Guo、N. Xue 和 S. Zafeiriou，“ArcFace: Additive Angular Margin Loss for Deep Face Recognition”，在*IEEE/CVF 计算机视觉和模式识别会议论文集*，2019 年，第 4690-4699 页。

[5] M. Breunig, H.-P. Kriegel、RT Ng 和 J. Sander，“Lof：识别基于密度的局部异常值”，在*2000 ACM SIGMOD 国际数据管理会议论文集*. ACM，2000 年，第 93-104 页。

[6] S. Ramaswamy、R. Rastogi 和 K. Shim，“从大数据集中挖掘异常值的有效算法”，载于 *2000 ACM SIGMOD 国际数据管理会议论文集*. ACM，2000 年，第 427-438 页。

[7] N. Harada、D. Niizumi、D. Takeuchi、Y. Ohishi、M. Yasuda 和 S. Saito，“ToyADMOS2：用于域转移条件下异常声音检测的微型机器操作声音的另一个数据集，” *arXiv 预印本 arXiv:2106.02369*, 2021.

[8] R. Tanabe、H. Purohit、K. Dohi、T. Endo、Y. Nikaido、T. Nakamura 和 Y. Kawaguchi，“MIMII DUE：用于工业机器故障调查和检查的声音数据集操作和环境条件的变化，” *在 arXiv 电子版：2006.05822, 1 –4*, 2021.

[9] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. 汤姆 é, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto、J. Mason、D. Ellis、E. Battenberg、S. Seyfarth、R. Yamamoto, K. Choi, viktorandreevichmorozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñu, F.-R。英石Ø ter、P. Friesch、A. Weiss、M. Vollrath 和 T. Kim，“librosa/librosa：0.8.0”，2020 年 7 月。[在线]。可用：https://doi.org/10.5281/zenodo.3955228

[10] O. Grisel, A. Mueller, Lars, A. Gramfort, G. Louppe, P. Prettenhofer、M. Blondel、V. Niculae、J. Nothman、A. Joly, TJ Fan, J. Vanderplas, manoj kumar, H. Qin, N. Hug, N. Varoquaux, L. Est è ve, R. Layton, JH Metzen, G. Lemaitre, A. Jalali, R. (Venkat) Raghav, J. Sch Ø 恩伯格, R. Yurchak、W. Li、C. Woolam、TD la Tour、K. Eren、J. du Boisberranger 和 Eustache，“scikit-learn/scikitlearn：scikit-learn 0.24.1”，2021 年 1 月。[在线]。可用：https://doi.org/10.5281/zenodo.4450597

表 3：Development Dataset 源域中 AUC 的谐波均值（%）

美国有线电视新闻网	探测器	合并	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门	全部的
基线（MNv2）			55.80	67.89	61.02	70.21	65.48	72.67	55.95	63.53
MNv2	洛夫	意思	72.68	63.48	82.18	78.31	75.46	82.02	74.31	74.99
		标准	62.11	60.64	69.09	63.88	59.10	76.47	88.67	67.31
	神经网络	意思	75.59	67.95	83.61	79.39	77.99	89.65	75.43	78.01
		标准	54.64	63.49	71.68	66.20	61.55	79.48	93.38	68.20
最惠国待遇	洛夫	意思	91.06	86.13	90.36	77.76	82.52	90.83	75.37	84.42
		标准	81.55	76.10	54.51	53.92	53.93	70.56	91.32	66.06
	神经网络	意思	89.37	81.50	85.59	79.52	83.37	91.97	71.31	82.73
		标准	75.53	73.36	51.69	53.76	53.53	72.43	95.87	64.97
我们最好的			91.06	86.13	90.36	77.76	82.52	90.83	95.87	87.42

表 4：Development Dataset 目标域中 AUC 的谐波平均值（%）

美国有线电视新闻网	探测器	合并	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门	全部的
基线（MNv2）			57.42	50.18	62.35	64.35	59.08	51.21	57.25	56.98
MNv2	洛夫	意思	64.48	53.77	64.12	77.88	62.15	67.22	67.22	64.58
		标准	59.65	53.50	62.41	59.07	59.53	60.07	73.46	60.63
	神经网络	意思	66.64	53.62	64.76	80.78	62.58	63.54	67.39	64.80
		标准	54.82	56.67	66.08	64.24	60.84	58.95	78.30	62.08
最惠国待遇	洛夫	意思	60.27	51.68	73.28	81.37	75.86	53.64	63.44	63.94
		标准	61.94	46.37	53.67	47.09	49.09	49.33	65.92	52.48
	神经网络	意思	70.54	54.08	72.67	84.80	74.39	67.07	63.10	68.34
		标准	61.62	42.55	53.85	47.96	47.55	49.49	78.67	52.59
我们最好的			70.54	54.08	72.67	84.80	74.39	67.07	78.67	70.50

表 5：Development Dataset 源域中部分 AUC 的谐波平均值（%）

美国有线电视新闻网	探测器	合并	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门	全部的
基线（MNv2）			58.64	51.87	65.79	61.45	59.24	59.50	52.17	58.01
MNv2	洛夫	意思	60.12	57.23	75.39	64.94	64.26	74.41	62.82	65.00
		标准	55.23	55.44	64.50	53.99	52.10	62.51	79.35	59.34
	神经网络	意思	63.48	57.64	75.80	64.67	65.34	78.68	64.07	66.43
		标准	53.85	54.17	67.70	54.27	54.40	71.31	82.70	61.02
最惠国待遇	洛夫	意思	78.25	65.36	79.66	67.67	66.50	83.05	59.41	70.48
		标准	63.27	58.72	50.91	51.07	51.93	58.35	79.19	57.81
	神经网络	意思	74.76	58.65	70.11	67.43	68.74	82.59	58.03	67.69
		标准	61.03	54.98	50.74	50.70	52.47	58.05	84.78	57.33
我们最好的			78.25	65.36	79.66	67.67	66.50	83.05	84.78	74.24

表 6：Development Dataset 目标域中部分 AUC 的谐波平均值（%）

美国有线电视新闻网	探测器	合并	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门	全部的
基线（MNv2）			54.44	51.38	60.84	57.48	55.73	53.17	53.23	55.03
MNv2	洛夫	意思	56.12	51.21	67.55	66.33	57.41	59.25	55.90	58.62
		标准	53.94	50.39	58.04	52.05	53.43	54.99	60.50	54.58
	神经网络	意思	57.74	51.35	67.82	65.57	58.06	58.98	56.13	58.93
		标准	52.39	51.22	64.82	52.67	54.19	56.80	64.41	56.18
最惠国待遇	洛夫	意思	56.91	52.45	64.81	69.04	62.38	50.55	57.13	58.40
		标准	52.64	48.64	49.32	49.22	51.41	51.97	56.80	51.30
	神经网络	意思	59.92	53.10	66.67	72.49	65.42	62.01	57.28	61.84
		标准	52.50	48.66	49.35	49.40	51.48	52.13	64.12	52.12
我们最好的			59.92	53.10	66.67	72.49	65.42	62.01	64.12	62.88

表 7：提交我们的系统

型号名称	特征提取器	异常检测器		合并方法
		源域	目标域	
森田 SECOM 任务 2 1 森田 SECOM 任务 2 2 森田 SECOM 任务 2 3	最惠国待遇	洛夫 神经网络 k-NN(阀门), LOF(否则)	洛夫 神经网络 神经网络	标准（阀门），平均值（否则）