从英语翻译成中文(简体) - www.onlinedoctranslator.com

域下互补异常检测器的集合状况技术报告

Jose A. Lopez, Georg Stemmer, Paulo Lopez-Meyer, Pradyumna S. Singh
胡安·德尔·霍约·昂蒂维罗斯，赫克托·库杜里埃

英特尔公司
{jose.a.lopez、georg.stemmer、paulo.lopez.meyer、pradyumna.s.singh}@intel.com
{juan.antonio.del.hoyo.ontiveros, hector.a.cordourier.maruri}@intel.com

抽象的

我们提交了对 DCASE2021 挑战任务 2 的提交，该任务旨在促进异常声音检测的研究。我们发现混合各种异常检测器的预测，而不是单独依赖众所周知的域适应技术，在域转移条件下为我们提供了最佳性能。我们提交的内容由两个自监督分类器模型、一个我们称为 NF-CDEE 的概率模型以及三者的集合组成。

索引词— DCASE，异常检测，域转移，机器状态监测，机器健康监测。

一、介绍

DCASE2021 挑战任务 2 涉及使用录音识别目标机器的异常行为 [1]。此任务与其他 DCASE 任务之间的主要区别在于它不受监督。因此，可用的训练数据仅包含来自正态分布的样本。增加这一挑战的另一个复杂因素是训练数据和测试数据的声学特性不同——这种情况被称为域转移，有一些已知的结果可以减少训练和测试数据之间的性能差距 [2 , 3, 4, 5, 6, 7, 8]。在我们的实验中，虽然我们认识到这些技术的潜力，但我们通常不会从单独使用这些方法中获得太多收益。

在我们提交的文件中，我们使用了两个自监督分类器，对部分 ID 进行分类，类似于几个团队在 DCASE2020 [9, 10, 11, 12, 13] 中采用的方法。对于第三个模型，我们引入了一个模型，该模型依赖于几个归一化流来估计输入 Mel 频谱图部分的条件密度，并使用它们的组合输出来产生异常分数 [14, 15, 16, 17, 18, 19, 20, 21, 22]。

在续集中，我们描述了每个模型、它是如何训练的、它的超参数以及它们各自的结果。为了正确看待结果，我们在表 1 和表 2 中包含了基线分数。本次挑战中使用的数据是 16 KHz、单通道、音频。有关更多详细信息，请参阅 [1, 23, 24]。

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
h-平均AUC	0.6249	0.6171	0.6324	0.6597	0.6192	0.6674	0.5341
h-均值 pAUC	0.5236	0.5381	0.5338	0.5276	0.5441	0.5594	0.5054

表 1：基线自动编码器分数

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
h-平均AUC	0.5604	0.5746	0.6156	0.6670	0.6189	0.5926	0.5651
h-均值 pAUC	0.5637	0.5161	0.6302	0.5916	0.5737	0.5600	0.5264

表 2：基线 MobileNetV2 分数

2. 架构

下面描述的第一个模型建立在 [9] 的工作基础上。特别是，编码在器网络已更新为使用 1D 卷积而不是 [9] 中的 2D。该模型的输入是带有或不带有 Mel 变换的频谱图。第二个模型建立在著名的 WaveNet 架构 [25] 之上，在扩张卷积之后添加了一个 x 向量 [26] 分类头——从某种意义上说，WaveNet 充当 xvector 组件的时间序列编码器。两种模型都经过训练以减少预测和部分 ID 之间的交叉熵损失。第三个模型与前两个模型的不同之处在于它是完全无监督的，并尝试学习一些以剩余 bin 为条件的 Mel 谱图 bin 的几种分布。由于不同的输入方式和学习方法，我们称这些方法是互补的。提供的最后一个系统是三个的集合。

我们所有的开发都是使用 PyTorch [27] 完成的，并且使用 nnAudio [28] 计算频谱图。第三个模型另外使用了 Pyro [29] 概率编程库。

2.1. XVector1D

第一个模型的架构的高级视图如表 3 所示。我们将附加边距 softmax 表示为 AMS [30]。

我们使用术语“标准化器”作为在将数据传递到网络的其余部分之前完成的预处理步骤。在大多数情况下，这只是一个禁用可学习参数的批处理规范层。通过这种方式，一旦运行的统计数据收敛，这个批处理规范将执行通常的频率归一化。然而，对于变速箱和 ToyCar，我们使用了 AutoDIAL 层 [4]。

玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
STFT	梅尔	STFT	梅尔	STFT	STFT	STFT
标准化者	标准化者	标准化者	标准化者	标准化者	标准化者	标准化者
编码器	编码器	编码器	编码器	编码器	编码器	编码器
x向量	x向量	x向量	x向量	x向量	x向量	x向量
AMS	AMS	AMS	AMS	AMS	AMS	AMS

表 3：XVector1D 高级架构

玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
自动拨号	批规范	自动拨号	自动拨号	自动拨号	自动拨号	自动拨号
C(128,192)	C(128,192)	C(128,192)	C(128,192)	C(128,192)	C(128,192)	C(128,192)
5 x C(192,192)	5 x C(192,192)	5 x C(192,192)	4 x C(192,192)	5 x C(192,192)	5 x C(192,192)	5 x C(192,192)
			C(192,90)			

表 4：编码器参数

该模型中使用的编码器使用内核大小为 3 的 1D 卷积和leaky-relu 激活。层数因机器而异，如表 4 所示 - 在此表中，我们使用“C”表示一维卷积。

这里使用的 x 向量分量与 [9] 中的基本相同，除了编码器的接口必须按预期进行调整以接受 1D 编码器输出。

2.1.1. 预处理

这个模型没有使用任何特殊的预处理或增强。STFT 和 Mel 谱图均采用对数。所有频谱图都是在频率最小值和最大值分别设置为 100 和 8000 赫兹的情况下计算的。

2.1.2. 培训和结果

该模型经过训练，可以使用分类交叉熵损失函数来预测部分 ID 元数据参数。我们发现频谱图参数对性能有很大影响。输入样本数、用于 FFT 的点数、跳跃长度等参数都会产生显着影响。我们通常使用 AdamW 优化器，默认学习率为1×10^{-3}和权重衰减设置为1×10^{-4}. 但是，我们对变速箱使用了具有默认学习率（并且没有权重衰减）的 ASGD。通常，训练损失使用 ASGD 收敛得更慢，但有时较慢的轨迹会花费更多的 epoch 接近 AUC 的最佳区域，这可以产生更好的结果。训练通常运行 300 个 epoch，使用来自开发和评估数据集的所有训练数据。最后，我们在训练期间使用来自最终 AMS 分类层之前的层的嵌入来计算平均嵌入。在测试时，平均嵌入用于计算到测试嵌入的余弦和马氏距离，作为异常分数的附加选项。结果如表5所示。

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
批量大小	128	64	128	64	128	128	64
输入样本	16384	16384	16384	98000	16384	16384	98000
不。梅尔斯	2048	128	2048	128	2048	2048	2048
不。快速傅立叶变换	4096	1024	4096	1024	4096	4096	4096
跳	80	512	512	80	512	512	512
得分	余弦	马哈拉诺比斯	软最大	马哈拉诺比斯	软最大	软最大	软最大
h-平均AUC	0.6702	0.7193	0.7171	0.8342	0.7799	0.7871	0.9032
h-均值 pAUC	0.6233	0.6772	0.7295	0.7443	0.6684	0.6728	0.7724

表 5：XVector1D 评分结果

2.2. WaveNet-XVector

我们探索了使用 WaveNet 模型直接处理音频样本。有关架构的详细信息，我们请读者参阅原始出版物 [25]。在原始论文中，作者解释说该模型可以很容易地适应分类任务，并且在他们的分类实验中，他们在扩张卷积之后添加了一个平均池化层，然后是“一些非因果卷积”。训练有两个损失项：一个是预测下一个样本，另一个是分类损失。我们遵循这个过程，因为我们使用平均池化层（内核大小为 10）并使用两个损失函数进行训练，但不是使用一些卷积，我们使用 x 向量组件，带有 AMS 顶层，与 XVector1D 一样模型。通过这种方式，可以将此模型视为使用纯音频编码器的 XVector1D 模型的变体。

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
输入过程	批规范	批规范	批规范	批规范	批规范	批规范	自动拨号
块	1	1	1	1	1	1	1
层	14	14	14	14	14	14	14
扩张通道	32	32	32	64	64	32	32
剩余通道	32	32	32	64	64	32	32
跳过通道	32	32	32	64	64	32	32

表 6：WaveNet 参数

2.2.1. 预处理

对于 Valve 和 ToyTrain，我们使用 Teager-Kaiser 能量算子对音频进行预处理 [31、32、33、34]。动机是，由于阀门噪声是稀疏和脉冲事件，Teager-Kaiser 算子提供的噪声抑制将提高阀门记录中的信噪比。尽管改善了 Valve 和 ToyTrain 的结果，但改善幅度不大。

2.2.2. 培训和结果

为了训练这个模型，我们使用了 Adamax 优化器和 200 个 epoch 的默认学习率。表 7 显示了该模型的性能。

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
批量大小	128	128	128	64	64	128	128
输入样本	16384	16384	16384	16384	16384	16384	16384
得分	软最大	软最大	软最大	软最大	软最大	软最大	软最大
h-平均AUC	0.5843	0.6641	0.8122	0.7156	0.7543	0.7184	0.7297
h-均值 pAUC	0.5629	0.5696	0.8025	0.5964	0.6506	0.6239	0.6206

表 7：WaveNet-XVector 评分结果

2.3. NF-CDEE

对于我们的第三个系统，我们尝试使用归一化流程对机器声音的 Mel 谱图的概率密度函数进行建模，对于单个机器。我们使用 Pyro [29] 概率编程库来开发这个模型。我们发现训练模型以拟合与 Mel bin 具有相同维度的分布有点不稳定。为了提高稳定性，我们改为估计几个条件密度并在单个模型中训练它们，最小化它们的负对数似然总和。我们认为这个模型是条件密度异常检测器的集合。因此，我们称这个模型为 NF-CDEE，因为它使用归一化流并且它是一个条件密度估计器集合。每个条件密度估计量都适合一个分布*n*-bin 以剩余 bin 为条件的输入频谱图的段。这减少了由于维度引起的不稳定性。参数*n*重叠量可由用户调整。对于这项工作，我们选择了*n* = 32 没有重叠。每个归一化流程都使用具有 16 个计数箱和默认隐藏层尺寸的单个条件样条——这些也是可调的，但在我们的实验中，它们并没有显着影响性能。

总而言之，每个估计器输出概率$磷(秒_{-种}|s_{-种C})$ 在哪里 *秒* 是一个维数等于数字的向量

梅尔・宾斯 *米* 由集合索引 *一世* = *{1, . . . , 米}*. *一种* 是一个 *n*-元素子集 *一世*，和 *一种c* 是它的补充 *我* -*一种*. 我们将正常状态的可能性定义为：

$$磷（正常）= \prod_{一世} 磷(秒_{一种}|S_{一种}, \theta) \quad (1)$$

在哪里 *一世* ∈ [1, . . . , *k*] 和 *克* 是用户提供的正整数——它是集成中估计器的数量。为了训练模型，我们最小化 *磷*（普通的）。因此，NF-CDEE 的输出是各个负对数似然的总和。

2.3.1. 培训和结果

为了训练这个模型，我们将输入音频转换为 256-bin Mel 频谱图，使用 8192 点 FFT 计算，跳跃长度为 512，并在传递到条件密度估计器之前应用频率归一化。每个模型都使用每种机器类型的开发（或评估）训练数据的所有部分进行训练——除了风扇，我们为每个部分训练了一个模型。为了进一步减少由标准化流行列式计算引起的训练不稳定性，我们在时间维度上取平均值。最后一步对于稳定合奏的训练很重要。如前所述，使用的损失函数是负对数似然的总和，这也作为异常分数。

对于优化器，我们使用了与 XVector1D 相同的优化器，带有梯度裁剪。在我们的实验中，这个模型通常需要训练大约 50 个 epoch。结果如表8所示。

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
批量大小	32	32	32	32	32	32	32
输入帧	192	192	192	192	192	192	192
米	256	256	256	256	256	256	256
n	32	32	32	32	32	32	32
克	8	8	8	8	8	8	8
得分	非本地语言	非本地语言	非本地语言	非本地语言	非本地语言	非本地语言	非本地语言
h-平均AUC	0.8657	0.7797	0.7866	0.8081	0.6993	0.7483	0.6130
h-均值 pAUC	0.7831	0.6031	0.6024	0.6513	0.5655	0.6054	0.5275

表 8：NF-CDEE 评分结果

2.4. 合奏

对于最后一个系统，我们通过首先标准化训练数据分数然后搜索凸组合网格来组合三个模型，类似于 [35]。结果如表9所示。

	玩具车	玩具火车	扇子	变速箱	泵	滑块	阀门
WaveNet权重	0.03	0.03	1.0	0.04	0.32	0.02	0
XVector1D 权重	0.06	0.55	0	0.61	0.68	0.52	1
NF-CDEE重量	0.91	0.42	0	0.35	0	0.46	0
h-平均AUC	0.8745	0.7756	0.8122	0.8613	0.7958	0.8287	0.9032
h-均值 pAUC	0.7837	0.7048	0.8025	0.7635	0.6790	0.6925	0.7724

表 9：合奏评分结果

3. 结论

我们已经概述了我们对 DCASE2021 挑战任务 2 的提交，其中包括训练和测试之间的领域转移分布。我们发现，域适应方法似乎对其他模态（尤其是视觉）表现良好，但对音频似乎效果不佳（至少在我们的实现中）。这种差异使 DCASE2021 挑战赛更具相关性，因为它突出了对音频社区的需求

为音频生成更有效的域适应方法。

在我们开发的模型中，我们发现 NF-CDEE 特别有前途，因为它是无监督的。在现实世界中，利用元数据并不总是可行的，即使可以这样做。此外，期望模型的集成特性在域转移情况下表现更好。展望未来，我们计划进一步开发此模型。

4. 参考资料

[1] Y. Kawaguchi、K. Imoto、Y. Koizumi、N. Harada、D. Niizumi、K. Dohi、R. Tanabe、H. Purohit 和 T. Endo，“关于 dcase 2021 挑战任务 2 的描述和讨论：用于在域转移条件下进行机器状态监测的无监督异常声音检测，” *arXiv 预印本 arXiv:2106.04492*, 2021.

[2] G. Wilson 和 DJ Cook，“无监督深度域适应调查”，2020 年。

[3] Y. Li、N. Wang、J. Shi、X. Hou 和 J. Liu，“用于实际领域适应的自适应批量归一化”，*模式识别*，卷。80，第 109-117 页，2018 年。

[4] FM Carlucci、L. Porzi、B. Caputo、E. Ricci 和 SR Bulò，“自动拨号：自动域对齐层”，2017 年。

[5] ——，“只需拨号：用于无监督域适应的域对齐层”，2017 年。

[6] M. Mancini、L. Porzi、SR Bulò、B. Caputo 和 E. Ricci，“通过发现潜在域促进域适应”，2018 年。

[7] J. Shen、Y. Qu、W. Zhang 和 Y. Yu，“Wasserstein 距离引导的领域适应表征学习”，2018 年。

[8] Y. Ganin 和 V. Lempitsky，“通过反向传播进行无监督域适应”，2015 年。

[9] JA Lopez、H. Lu、P. Lopez-Meyer、L. Nachman、G. Stemmer 和 J. Huang，“异常检测的说话人识别方法”，在 *声场景和事件的检测和分类 2020 Workshop (DCASE2020) 论文集*，日本东京，2020 年 11 月，第 96-99 页。

[10] R. Giri、SV Tenneti、F. Cheng、K. Helwani、U. Isik 和 A. Krishnaswamy，“用于检测异常声音的自监督分类”，在 *声场景和事件的检测和分类 2020 Workshop (DCASE2020) 论文集*，日本东京，2020 年 11 月，第 46-50 页。

[11] T. Inoue、P. Vinayavekhin、S. Morikuni、S. Wang，T. Hoang Trong、D. Wood、M. Tatsubori 和 R. Tachibana，“使用分类置信度检测机器状态监测的异常声音”，在 *声场景和事件的检测和分类 2020 Workshop (DCASE2020) 论文集*，日本东京，2020 年 11 月，第 66-70 页。

[12] P. Primus、V. Haunschmid、P. Praher 和 G. Widmer，“异常声音检测作为一个简单的二元分类问题，仔细选择代理异常值示例，”

在 *声场景和事件的检测和分类 2020 Workshop (DCASE2020) 论文集*，日本东京，2020 年 11 月，第 170-174 页。

[13] Q. Zhou，“用于 dcase 2020 任务的基于 Arcface 的声音移动网络 2，”DCASE2020 挑战赛，技术。众议员，2020 年 7 月。

[14] EG Tabak 和 CV Turner，“非参数密度估计算法系列，”*纯数学与应用数学通讯*，卷。66，没有。2，第 145-164 页。

[15] DJ Rezende 和 S. Mohamed，“标准化流的变分推理”，2016 年。

[16] I. Kobyzev、S. Prince 和 M. Brubaker，“规范化流程：当前方法的介绍和回顾”，*IEEE 模式分析和机器智能汇刊*，第 2020 年 1-1 日。

[17] G. Papamakarios、E. Nalisnick、DJ Rezende、S. Mohamed 和 B. Lakshminarayanan，“概率建模和推理的标准化流”，2021 年。

[18] C. Durkan、A. Bekasov、I. Murray 和 G. Papamakarios，“神经样条流”，2019 年。

[19] HM Dolatabadi、S. Erfani 和 C. Leckie，“使用线性有理样条的可逆生成建模”，2020 年。

[20] L. Dinh、D. Krueger 和 Y. Bengio，“Nice：非线性独立分量估计”，2015 年。

[21] L. Dinh、J. Sohl-Dickstein 和 S. Bengio，“使用真实 nvp 的密度估计”，2017 年。

[22] D. Ha、A. Dai 和 QV Le，“超网络”，2016 年。

[23] 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.

[24] 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.

[25] A. van den Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior 和 K. Kavukcuoglu，“Wavenet：原始音频的生成模型”，2016 年。

[26] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，“X 向量：用于说话人识别的稳健 dnn 嵌入”，在 *2018 IEEE 声学、语音和信号处理国际会议 (ICASSP)*. IEEE，2018 年，第 5329-5333 页。

[27] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga、A. Desmaison、A. Kopf、E. Yang、Z. DeVito、M. Raison、A. Tejani、S. Chilamkurthy、B. Steiner、L. Fang、J. Bai 和 S. Chintala，“Pytorch：一种命令式风格的高性能深度学习库”，在 *神经信息处理系统的进展 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, 和 R. Garnett, Eds. Curran Associates, Inc.，2019 年，第 8024-8035 页。

[28] KW Cheuk、H. Anderson、K. Agres 和 D. Herremans，“nnaudio：使用一维卷积神经网络的实时 gpu 音频到频谱图转换工具箱”，2020 年。

[29] E. Bingham、JP Chen、M. Jankowiak、F. Obermeyer、N. Pradhan、T. Karaletsos、R. Singh、PA Szerlip、P. Horsfall 和 ND Goodman，“Pyro：深度通用概率规划”，*J. 马赫。学习。水库*，卷。20，第 28:1–28:6，2019 年。

[30] F. Wang、J. Cheng、W. Liu 和 H. Liu，“用于人脸验证的附加边距 softmax，”*IEEE 信号处理快报*，卷。25，没有。7，第。926-930，2018 年 7 月。

[31] P. Maragos、J. Kaiser 和 T. Quatieri，“信号调制中的能量分离与语音分析的应用”，*IEEE 信号处理汇刊*，卷。41，没有。10，第 3024-3051 页，1993。

[32] P. Maragos、JF Kaiser 和 TF Quatieri，“使用能量算子进行幅度和频率解调”，*IEEE 信号处理汇刊*，卷。41，没有。4，第 1532-1550 页，1993 年 4 月。

[33] A. Georgogiannis 和 V. Digalakis，“在嘈杂环境中使用基于非线性能量的特征的语音情感识别”，在 *2012 第 20 届欧洲信号处理会议 (EUSIPCO) 论文集*，2012，第 2045-2049 页。

[34] H. Li、H. Zheng 和 L. Tang，“基于teager-huang 变换的齿轮故障检测”，*国际旋转机械杂志*，卷。2010 年，第 1-9 页，2010 年。

[35] P. Daniluk、M. Gozdziewski、S. Kapka 和 M. Kosmider，“基于自动编码器的异常检测系统集合”，DCASE2020 挑战赛，技术。众议员，2020 年 7 月。