声音事件检测项目报告

520030910342 柳纪宇

1. 概述

声音事件检测任务旨在使机器能够通过输入的声学特征识别出对应的场景以及其起始、结束时间，在智能机器人、智能交通、自动辅助驾驶等领域中有很强的应用价值。在本次实验中，我了解了声音事件检测任务的基本流程、模型设计原理与任务难点，并使用 CRNN 模型完成了 DCASE18 数据集上的训练与检测任务。

2. 基本流程与模型介绍

2.1. 基本流程介绍

声音事件检测任务的基本流程涵盖了以下几个关键步骤，以确保准确地识别声音事件：首先，从输入的音频信号中提取声学特征，包括时域特征和频域特征。接着，对该信号进行重采样，使其具有固定的采样率。然后，将信号转换为梅尔频谱并转化为分贝类型，这样我们就能够获得从声音文件中提取出的固定频率的声学特征。

接下来，我们使用这些特征来训练深度学习模型，例如 CRNN（卷积循环神经网络）模型。CRNN 模型结合了卷积神经网络（CNN）和循环神经网络（RNN）的优点，能够捕捉时序信息和频域特征，从而提高声音事件的识别准确性。在此过程中，我们将提取的特征作为 CRNN 模型的输入。这些特征经过 CNN 层后，生成具有声音文件时间长度的特征图。然后，将该特征图作为 RNN（或其变体 LSTM 与 GRU）的输入，得到预测向量。在预测向量中，数值大于一定阈值的位置将被视为预测的事件输出。

为了优化模型的性能，我们计算预测向量与真实值之间的预测误差（本实验中使用二元交叉熵误差），并使用梯度回传算法来更新模型参数。这样，模型将逐步调整以更准确地预测声音事件。

2.2. 评估指标

该任务的评估方式有分割层面 (segment-level)、事件层面 (event-level) 和标注层面 (tagging-level) 三种，在每种层面上分别有准确率 (precision)、召回率 (recall) 和 F1-Score 三种。

在 segment-level 评估中，重点是评估声音事件检测在单个分段或时间间隔的准确性。每个分段通常对应于一小段音频，例如几秒钟。在这个层级上的评估指标评估系统在每个分段中检测特定声音事件的存在或缺失的准确程度。

event-level 评估考虑声音事件检测在完整事件的级别上的性能表现，一个事件可能跨越多个分段或时间间隔。该层面上的评估指标为模型预测得到的事件的开始时间和结束时间是否在标签中开始时间和结束时间的一定范围之内。

tagging-level 评估侧重于评估声音事件的分类或标记的准确性。它评估系统在没有明确考虑时间边界或持续时间的情况下，对音频样本或分段正确分配事件标签的能力。

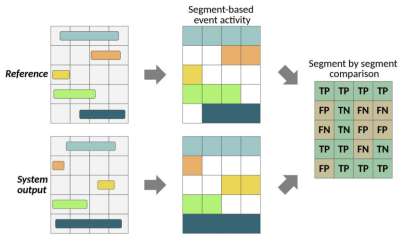

图 1: 分割层面的 *F1 Score* 示意图

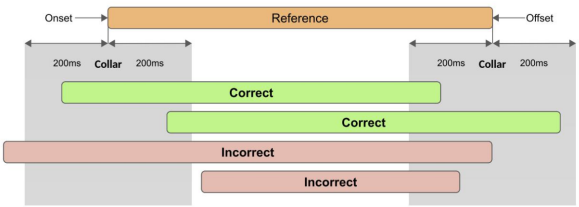

图 2: 事件层面的 *F1 Score* 示意图

2.3. Baseline 模型

本部分我将简单介绍课件中给出的 Baseline 模型的设计原理。

Baseline 模型采用了 CNN 加双向 GRU 的基本结构（如图3所示）。其中 CNN 由五个由卷积层、BatchNorm 层、ReLU 层和最大值池化层构成的卷积块组成。卷积层和池化层保证了特征的提取以及感受野大小的协调，BatchNorm 层能够使得模型训练更加快速且稳定。值得注意的是，在第二个卷积块之后所有的最大值池化层中的 kernel 大小都从 (2, 2) 更改成了 (1, 2)，这是因为输入的声学特征的长（时间长度）一般会比宽（梅尔频率维数）大很多，这样设计避免了输出特征的宽长比变得太小。

该模型的流程为将声学特征输入 CNN 中得到深层特征，接着将该深层特征输入双向 GRU 模型得到概率张量，接着再通过线性层和 Sigmoid 激活函数，得到范围在 0 到 1 之间的概率向量。

2.4. 任务难点

一般情况下声音事件检测任务有两种数据标注方式：强标注（strongly-labeled）和弱标注（weakly-labeled）。其中强标注数据不仅包括事件的类别，还包含开始时间与结束时间；而弱标注数据仅包括事件类别。尽管强标注数据对训练有着更好的监督效果，但由于一些事件的持续时间很长、重复出现次数多等特点，强标注数据对标注人力的要求是巨大的，因此主流声音事件检测任务主要使用的是弱标注数据集。接下来我将介绍弱标注数据下声音事件检测任务的几个难点。

- 尽管使用弱标注方法能够产生出更多的标注数据，但标注的精度也受到了相应的影响。为了使模型能够更好地学习到声音事件的开始结束时间，该流程引入了 RNN 及其变体作为网络结构的一部分，这使得模型能够较好地学习到输入特征的时序信息。
- 声音事件检测任务的另一大难点是数据稀疏：在弱标注数据集中有标注的信息可能远少于无标注信息。过去的研究者提出了迁移学习方法来应对这一难题，通过在大数据上训练提取音频特征的神经网络模型以解决分布外泛化问题。
- 日常生活中发生的事件通常是并行的，这意味着同一个音频文件中可能对应着多个事件。这给声音事件检测任务增添了更大的复杂性和不确定性。

2.5. Baseline 实验结果

使用如上介绍的 baseline 模型进行 100 个 epoch 的训练，得到的结果如表1所示：

baseline	f_measure	precision	recall	mAP
event	0.1339	0.108636	0.217249	
segment	0.584559	0.589875	0.602564	0.6619
tagging	0.641991	0.676215	0.622777	

表 1: *Baseline* 模型各指标评测结果

3. 性能改进

3.1. 数据增强

经过调查 [1][2]，我在原有数据集的基础上增加了五种数据增强方式，它们分别是：时移 (time shift)、时间遮挡 (time mask)、增加均匀噪音 (add uniform noise)、增加高斯噪音 (add gaussian noise)、音频融合 (mix audio)。以下是对这几种数据增强方式的详细介绍：

- 时移：随机将原特征向右移动从 0 到最大时长之间的一个随机整数，左侧的空余位置用 0 来填充，将输出时移后的特征与原标签作为输出。
- 时间遮挡：以某固定遮挡率对原特征进行遮挡（遮挡结果设置为 0），将遮挡后的特征与原标签作为输出。
- 增加均匀噪音：从最小幅值和最大幅值之间均匀采样一个值作为均匀分布噪音的幅值，将该噪音与原特征相加后的结果与原标签作为输出。

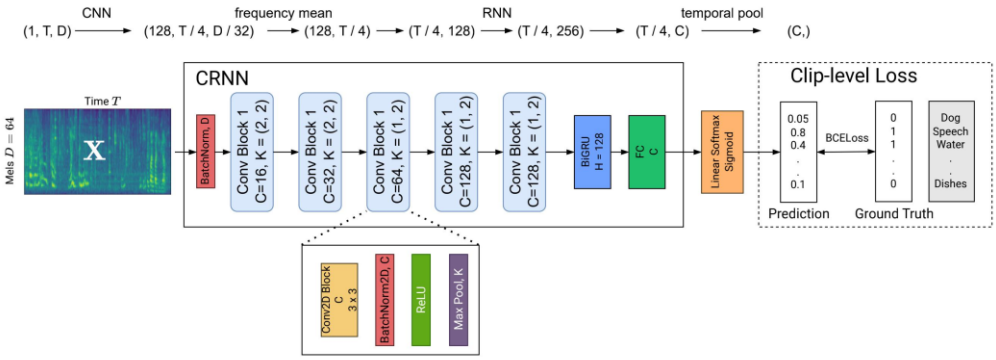

图 3: *Baseline* 模型结构示意图

- 增加高斯噪音：以设定均值和标准差从高斯分布中采样与原特征形状相同的噪音，将该噪音与原特征相加后的结果与原标签作为输出。

- 音频融合：随机取一组音频特征与标签与现有特征与标签进行融合。本实验中对特征的融合方式为加权求和，对标签的融合方式有逻辑或 (or)、逻辑与 (and) 和加权求和 (add) 三种方式组成。

首先，我对每种数据增强方式分别作用于数据集的效果进行了检验（注：由于增加均匀噪音与增加高斯噪音的效果区别不大，因此在实施时我将两种增强方式合并在一起执行），在原有数据集的基础上加入了等量增强后的数据，得到的结果见表2（各项评测指标的最佳结果由黑色加粗字体标出）。

分析这一实验结果，可得到如下结论：

- 时间遮挡：与 Baseline 相比，时间掩码对 Segment 层面和 Tagging 层面的各项指标分数均有所提高，这表明，通过在时间域上随机屏蔽部分数据，可以增加模型的泛化能力，对某些任务有正面影响。然而在 Event 层面的各项指标均不如 Baseline，这应该是因为遮挡操作对一个事件的起始位置和结束位置的预测产生了负面影响。

- 时移：相对于 Baseline，时移在除 Event 层面之外的几乎所有指标上都有所提高，且在 mAP 这一项上达到了最佳。这说明，在时间轴上移动音频信号可以帮助模型学习到更稳健的特征，从而提高模型性能。但与时间遮挡相同，该增强方式会影响对事件起始位置和结束位置的预测效果。

- 增加噪音：在所有任务中，添加噪声的各项指标分数都高于 Baseline。这表明，通过给音频信号添加噪声，可以很好地提高模型的泛化性能和鲁棒性。

- 音频融合：使用 or 操作处理标签在 Event 层面明显劣于 Baseline，而在其他层面与 Baseline 相差不大，推测可能是因为在特征上进行加权求和使得模型难以检测到融合特征中的全部事件。使用加权求和处理标签总体优于 or 操作处理，但仍然有着与 or 操作类似的问题。使用 and 操作处理标签在精度上表现地很好，但其他指标（特别是召回率）都明显劣于 Baseline，这是因为 and 操作既减少了模型预测错误事件的概率，也导致了部分标签信息的丢失。

在完成上述实验后，我发现在音频融合这一增强方式中，音频特征和数据结合的方式可能对增强的效果有很大影响，故我尝试将融合的比例分别调整为 1:1（原增强所用比例）、7:3、4:1、9:1 进行实验，得到的结果如表3所示（各项指标的最佳结果有加粗黑色字体标出）。

综合来看，4:1 的融合比例下模型预测效果最佳，而 7:3 和 9:1 的融合比例在 Tagging 层面的精度和召回率上分别有着较好的表现。

Augmentation		Baseline	Time Mask	Time Shift	Add Noise	Mix OR	Mix Add	Mix And
Event	F Measure	**0.1339**	0.120103	0.0851493	0.125365	0.0697821	0.0844752	0.0991372
	Precision	**0.108636**	0.09643	0.0695245	0.10312	0.0570768	0.0662296	0.0812599
	Recall	**0.217249**	0.193424	0.14268	0.197794	0.127573	0.149259	0.138695
Segment	F Measure	0.584559	0.594437	0.604522	**0.605868**	0.575428	0.585811	0.529658
	Precision	0.589875	0.598335	0.59602	**0.608409**	0.557438	0.588818	0.599538
	Recall	0.602564	0.617395	0.630566	**0.628163**	0.609147	0.605159	0.495937
Tagging	F Measure	0.641991	0.658449	0.651677	**0.661911**	0.622279	0.639974	0.546259
	Precision	0.676215	0.675965	0.671557	0.686885	0.687435	0.685461	**0.715666**
	Recall	0.622777	**0.654405**	0.64436	0.645714	0.580116	0.609142	0.475875
	mAP	0.6619	0.6724	**0.6841**	0.6813	0.6585	0.6726	0.6169

表 2: *Baseline* 及各增强方式单独作用下的评测指标汇总表

Ratio	Metric	F Measure	Precision	Recall
1:1	Event Based	0.0844752	0.0662296	0.149259
	Segment Based	0.585811	0.588818	0.605159
	Tagging Based	0.639974	0.685461	0.609142
	mAP	0.6726412219796298		
7:3	Event Based	0.0733281	0.0586526	0.125899
	Segment Based	0.579161	0.584491	0.590752
	Tagging Based	0.614341	**0.71716**	0.547551
	mAP	0.6682694433627292		
4:1	Event Based	0.0891957	0.0734188	0.145141
	Segment Based	0.608342	0.609227	0.627732
	Tagging Based	**0.66412**	0.711913	0.628877
	mAP	**0.6841375202718292**		
9:1	Event Based	0.112525	0.0932575	0.176645
	Segment Based	0.588057	0.588562	0.613648
	Tagging Based	0.653287	0.675744	**0.640899**
	mAP	0.6760701864203567		

表 3: 音频融合不同融合比例下的各评测指标汇总表

在完成了各数据增强方式单独作用的实验中，我尝试了将各增强方式组合使用。我尝试的组合一共有：

1. 依次使用上述所有增强方式 (Apply All)
2. 从所有增强方式中随机采样一种应用到数据上 (Random Apply)
3. 将时移和增加噪音一起应用 (Shift and Noise)
4. 随机采用时移和增加噪音中的一种 (Shift or Noise)
5. 随机增加噪音和音频融合中的一种，融合比例设置为 4:1 (Noise or Mix(4:1))
6. 随机采用时移和音频融合中的一种，融合比例设置为 1:1 (Shift + Mix (1:1))
7. 随机采用时移和音频融合中的一种，融合比例设置为 4:1 (Shift + Mix (4:1))
8. 随机采用时移、增加噪音和音频融合（比例为 4:1）中的一种 (Shift or Noise or Mix (4:1))

得到的实验结果汇总可见表4（各评测指标的最佳结果由黑色加粗字体标出）。

分析实验结果，我们发现：

- 对每组数据依次使用所有增强方式能够在 Event 层面上获得很好的预测结果，但每种单独作用的数据增强在 Event 层面上的评测分数均不如 Baseline。推测可能是因为多种数据增强一起使用使得数据的分布更为多样，因此模型在这些数据上能够学习到更具泛化性的事件层面的信息。

- 在与时移的组合下，1:1 融合比例的音频融合比 4:1 融合比例的音频融合效果更好。推测原因可能是：音频融合主要是在特征级别上进行操作，而时移则是在时间轴上进行操作：在融合比例为 4:1 时二者之间可能出现冲突，故预测表现较差；而融合比例为 1:1 时特征级别上的操作得到了加强，从而与时间轴上的操作形成了一定程度上的互补，故预测表现较好。

- 在实践中，对同一组数据应用多种增强方法可能并不总是提升模型的最终预测效果。这种现象的原因可能是多种增强方法之间的相互冲突和对数据分布的潜在影响。具体来说：第一，各种增强方式对数据的作用可能同时存在着增强和抑制；第二，过多的增强可能导致增强后的数据和原数据的分布差距较大。

3.2. 调节隐层大小

在结束数据增强实验后，我还从模型结构的角度尝试改进模型性能。首先，我修改了 CRNN 模型的隐层大小，将其分别设置为 128（baseline）、256、512 进行实验，得到的结果如表5所示。

从实验结果可以看出，适当地增大 CRNN 模型的隐层大小对声音事件识别任务的性能有较大提升。

3.3. 调节 ConvBlock 个数

此外，我还尝试了修改模型中的 ConvBlock 个数，将其设置为 3、4、5 进行实验，得到的结果如表6所示。

从实验结果我们可以得知，随着 ConvBlock 个数的增加（即模型深度的增加），预测准确度有着较为显著的提升。

3.4. 其他尝试

根据之前的实验结果，我发现时移和音频融合这两种数据增强方式对模型性能的提升效果最好。同时，将隐层大小设置为 256、将 ConvBlock 个数设置为 5 时模型的性能达到最佳。因此，我尝试了将这两种数据增强方式与上述模型结构结合起来进行训练，以期达到最佳的模型性能。

我得到的结果如表7所示，由结果我们可以看出，尽管时移、音频融合和上述模型结构都能显著提升模型性能，但将二者结合在一起时得到的结果却没有单独使用隐层大小为 256、ConvBlock 个数为 5 的 CRNN 模型好，我认为可能的原因为隐层大小增大后模型需要的训练数据也随之增多，在数据增强将原样本分布数据稀释后应该需要更多的数据来训练模型，因此在相同 epoch 数量下使用数据增强反而会使性能下降。

4. 结论

在本次实验中，我实现了 CRNN 模型下的声音事件检测任务，了解了模型设计原理和任务难点。除此之外，我还从数据增强、模型深度调整、模型参数调整等角度进行了一系列尝试。最终，在隐层大小为 256 的条件下，我在测试集上达到了最佳 mAP 0.6967。

5. 参考文献

[1] https://github.com/alibugra/audio-data-augmentation

[2] https://github.com/iver56/audiomentations

Method	Metric	F Measure	Precision	Recall
Apply All	Event Based	**0.153593**	**0.133769**	**0.213549**
	Segment Based	0.557014	**0.627784**	0.523318
	Tagging Based	0.620948	0.673238	0.586806
	mAP	0.6388601775323905		
Shift and Noise	Event Based	0.0782982	0.0640405	0.13399
	Segment Based	0.584473	0.573611	0.608549
	Tagging Based	0.627313	0.667325	0.602683
	mAP	0.6660859353311865		
Noise or mix (4:1)	Event Based	0.0994871	0.0836719	0.157978
	Segment Based	0.598251	0.600685	0.623575
	Tagging Based	0.654715	**0.708526**	0.621144
	mAP	0.6814784912940528		
Shift or Noise	Event Based	0.136668	0.112623	0.21347
	Segment Based	0.604207	0.597212	**0.628492**
	Tagging Based	0.657633	0.675183	**0.650415**
	mAP	0.681352309371971		
Shift or Mix(1:1)	Event Based	0.101391	0.0816658	0.174432
	Segment Based	**0.597872**	0.594287	0.619031
	Tagging Based	**0.66271**	0.688536	0.649331
	mAP	**0.6849664544433061**		
Shift or Mix (4:1)	Event Based	0.0936613	0.0746988	0.165081
	Segment Based	0.585892	0.576372	0.614503
	Tagging Based	0.639242	0.674452	0.61788
	mAP	0.671853668374175		
Shift or Noise or mix (4:1)	Event Based	0.094141	0.0777778	0.151857
	Segment Based	0.586691	0.582004	0.614571
	Tagging Based	0.635262	0.665327	0.61653
	mAP	0.6727734857991932		

表 4: 各组合增强方式下的各评测指标汇总表

Method	Metric	F Measure	Precision	Recall
Hidden Size: 128	event_based	0.1339	0.108636	0.217249
	segment_based	0.584559	0.589875	0.602564
	tagging_based	0.641991	0.676215	0.622777
	mAP	0.6618630299394379		
Hidden Size: 256	event_based	0.124518	0.109092	0.187535
	segment_based	0.602453	0.60361	0.613649
	tagging_based	0.654427	0.68627	0.637942
	mAP	0.6966858171076014		
Hidden Size: 512	event_based	0.124603	0.107698	0.179885
	segment_based	0.603778	0.618491	0.600356
	tagging_based	0.650188	0.68679	0.629275
	mAP	0.6833306338740488		

表 5: 各隐层大小下的各评测指标汇总表

Method	Metric	F Measure	Precision	Recall
Number of ConvBlock: 3	event_based	0.0177378	0.0114939	0.0568815
	segment_based	0.546976	0.529588	0.577568
	tagging_based	0.587319	0.630627	0.56633
	mAP	0.620146207747142		
Number of ConvBlock: 4	event_based	0.04729	0.0345662	0.114788
	segment_based	0.574947	0.55286	0.617011
	tagging_based	0.626362	0.642049	0.620015
	mAP	0.6541549222170989		
Number of ConvBlock: 5	event_based	0.1339	0.108636	0.217249
	segment_based	0.584559	0.589875	0.602564
	tagging_based	0.641991	0.676215	0.622777
	mAP	0.6618630299394379		

表 6: 各 *ConvBlock* 个数下的各评测指标汇总表

Method	Metric	F Measure	Precision	Recall
Hidden_256 + Shift	event_based	0.0883522	0.0732209	0.146085
	segment_based	0.587492	0.580431	0.607838
	tagging_based	0.653212	0.67038	0.648054
	mAP	0.689249092953107		
Hidden_256 + Mix(0.5)	event_based	0.0972182	0.0783396	0.167591
	segment_based	0.603415	0.602374	0.620674
	tagging_based	0.663372	0.708225	0.632543
	mAP	0.6917667044105702		
Hidden_256 + Mix(0.8)	event_based	0.112969	0.0957817	0.180067
	segment_based	0.600937	0.607143	0.607563
	tagging_based	0.667446	0.711741	0.637716
	mAP	0.6839018876974882		
Hidden_256 + Shift or Mix(0.5)	event_based	0.0903756	0.0740758	0.156798
	segment_based	0.597082	0.599773	0.607002
	tagging_based	0.661673	0.6927	0.643778
	mAP	0.6951398305527002		

表 7: 时移、音频融合在隐层大小为 *256* 时的各评测指标汇总表