北京交通大学

信息网络综合专题研究课

应用层论文阅读笔记

学院： 电子信息工程学院

班级： 通信 1903 班

指导教师： 陈一帅 郭宇春

姓名学号：关博予 19211007

北京交通大学

2022 年 5 月

一、文献信息

1. 论文题目：SpecMix : A Mixed Sample Data Augmentation method for Training with Time-Frequency Domain Features
2. 作者：Gwantae Kim, David K. Han, Hanseok Ko
3. 发表途径：Interspeech 2021
4. 时间：2021 年

二、问题背景

近年来，随着深度学习技术不断发展，它在音频处理上的应用取得了很多成果，例如语音合成、语音识别、音频分类等应用场景。很多研究人员都将工作重点聚焦在为这些任务场景设计更好的网络结构上。这些研究固然很有必要，但是我们知道，对于深度学习来说，充足的训练数据是不可或缺的，否则很容易造成过拟合。在解决扩充训练数据量的问题上，数据增强是必不可少的一项技术。数据增强可以在数据集有限的前提下，对其进行适当处理，达到扩充数据量，改善训练结果的目的。

在音频数据的数据增强问题上，有两种主要方法：一是时域波形的数据增强，二是时频域特征的数据增强。时域波形的数据增强方式包括加入噪声、改变音调、改变速度等。时频域特征包括频谱图、Mel 频谱图、MFCC 等。由于时频域特征是二维的，可以用二维图像表示，因此计算机视觉领域的数据增强方式可以应用于时频域特征的数据增强。但是，因为时频域特征本质上是关于音频频率的时间序列，所以采用一般的图像数据增强方式可能会导致音频信息的丢失，在实际应用中需要更适用于时频域特征的数据增强方式。论文作者正是为了解决这一问题，提出了名为 SpecMix 的数据增强方式，采用了一系列专门为时频域特征设计的数据增强策略。这一方式可以应用到 ResNet 等网络结构中，用于解决语音增强、音频分类等任务。

三、思路方法

在不同的音频处理任务中，具体所需要的数据增强方式可能会有所不同。但是大体上，各类任务中的数据增强都可以通过组合两个不同的训练样本，生成一个新的样本来实现。在论文中，作者正是根据这种思路，进行了针对不同任务的数据增强算法设计。

1. 音频分类任务中的算法

进行分类任务需要有数据和其标签。设$x \in \mathbb{R}^{F\times T\times C}$和 y 分别表示数据及其标签。其中 F 表示频率维度的长度，T 表示时间维度的长度，C 表示时频域特征的数量。SpecMix 算法的目标将两个训练样本(x_A, y_A)和(x_B, y_B)组合起来，生成一个新的样本$(\tilde{x}, \tilde{y})$。论文进行组合的方式如下：

$$\tilde{x} = \mathrm{M} \odot x_A + (1 - \mathrm{M}) \odot x_B \qquad (1)$$

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \qquad (2)$$

其中$\mathrm{M} \in \{0,1\}^{F\times T}$表示掩码矩阵，用于进行两个时频域特征图像的组合。⊙是按元素的乘法（区别于一般的矩阵乘法）。数据标签的组合比 λ 是组合数据$\tilde{x}$中x_A像素点所占的比例。具体在每一次训练迭代中，可以根据上式在两个 mini-batch 中选择数据组合成为新的混合样本。

2. 语音增强任务中的算法

进行语音增强时，需要从噪声背景下提取有用的语音信息。训练时数据中需要有含噪声信号和清晰信号。设$x \in \mathbb{R}^{F\times T\times C}$与$z \in \mathbb{R}^{F\times T\times C}$分别表示含噪声信号的时频域特征和清晰信号的时频域特征。数据增强的目

标是组合两个训练样本(x_A, z_A)和(x_B, z_B)生成一个新的样本$(\tilde{x}, \tilde{z})$。此时的组合定义为：

$$\tilde{x} = \mathrm{M} \odot x_A + (1 - \mathrm{M}) \odot x_B \quad (3)$$

$$\tilde{z} = \mathrm{M} \odot z_A + (1 - \mathrm{M}) \odot z_B \quad (4)$$

其中 M 与⊙与音频分类任务中的含义一致。具体来说，在每一次训练迭代中，同样可以在两个 mini-batch 中选择数据组合生成新的混合样本。

3. 掩蔽

在实施 SpecMix 算法时，通过掩码矩阵 M 对两组数据进行频率与时间掩蔽生成组合数据。通过频率掩蔽可以生成最多三个不同的频率段，通过时间掩蔽也可以生成最多三个不同的时间段。频率段与时间段的长度都可以由操作者根据实际需要来定义。生成组合数据的算法示意图如图 1。

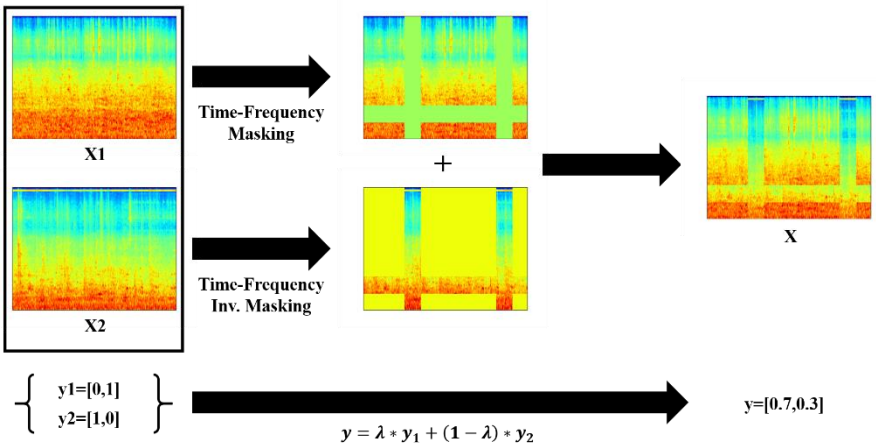

图 1　SpecMix 算法流程示意

具体来说，在实行频率掩蔽时，生成频率段的数量f_{times}可以从 0 到 3 中随机选取整数值。在每个频率段的选择上，起始频率f_{start}可以从 0 到 F 中随机选择。然后利用公式$f_{end} = f_{start} + \gamma F$得出终止频率。$\gamma$可以由操作者从 0 到 1 中选取。这一过程的重复次数为f_{times}。时间掩蔽的过程与频率掩蔽的过程相似。

四、实验及结论

在提出了数据增强的具体算法后，结合具体的网络与任务进行实验是必不可少的操作流程。作者分别就声音场景分类、事件分类与语音增强三个任务进行了实验，并且与其他音频数据增强算法得到的结果进行了对比。

1. 所用模型

1.1 场景分类与事件分类

对于场景分类与事件分类这两项任务，作者采用的时频域特征是 Mel 频谱图及其一阶差分和二阶差分特征。音频的采样率是 44.1kHz，进行 FFT 变换的位数为 2048，中间跳跃点数为 1024，Mel 滤波器组数量为 128。因此输入数据的形状为$[F, T, C]$=$[128, T, 3]$，其中 F 是频率维度的长度，T 是时间维度的长度，C 是音频的通道数。

在网络模型的选取上，作者采用 ResNet-101 用于分类任务。训练时选择 Adam 优化器与交叉熵损失函数，训练批次大小（batch size）为 32。此外，作者还采取了学习率衰减策略，学习率从10^{-3}衰减至10^{-7}。

除 ResNet-101 外，作者还选取了 M. D. McDonnell 等人于 ICASSP 2020 提出的最新音频场景分类模型进行测试，来检验 SpecMix 算法的泛化性能。

1.2 语音增强

在语音增强任务中，作者首先将音频的时域波形进行调整，使得波形长度均为 32768。接下来，作者采用频谱图作为时频域特征。音频的采样率是 16kHz，进行 FFT 变换的位数为 512，中间跳跃点数为 256。因此输入数据的形状为[F, T, C] =[256, T, 2]，其中 F 是频率维度的长度，T 是时间维度的长度，C 是音频的通道数。C 的值为 2 是因为分别采用了频谱图的实部与虚部作为输入的两个通道。

为了进行训练，作者构建了 U-Net 语音增强模型，如图 2 所示。这一模型中包含 8 个编码器层、1 个中间层、8 个解码器层和 1 个卷积层。训练时选择 Adam 优化器与均方误差损失函数，训练批次大小为 6。作者还采取了10^{-2}至10^{-5}的学习率衰减。

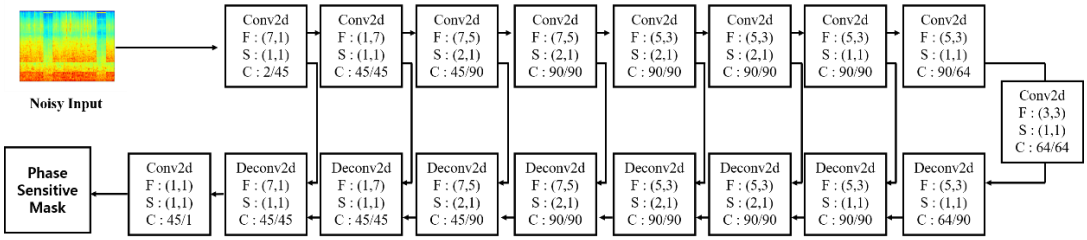

图 2　用于语音增强的 U-Net 模型

2. 实验设计

2.1 场景分类实验

在此实验中，作者使用了 TAU Urban Acoustic Scenes 2020 Mobile benchmark 数据集。该数据集包含来自 12 个欧洲城市的 10 个不同声音场景的录音数据，由 4 种不同的录音设备录制而成。其中有 13965 个片段用于训练，2970 个片段用于测试。评估时使用准确率作为指标。

为了进行对比，作者采用了无数据增强的处理方式，以及 Mixup、Cutmix 和 SpecAugment 三种数据增强算法。这三种算法都是受传统图像数据增强算法的启发，进行音频数据增强。在 ResNet-101 上进行测试的结果如表 1 所示。结果表明，相较于其他方式，作者提出的 SpecMix 算法获得的准确率最高（62.13%），比 Mixup、Cutmix 和 SpecAugment 三种数据增强算法分别高出 3.98%、2.59%和 4.45%。值得注意的是，这三种算法的结果甚至不如无数据增强处理的结果。因此，作者认为基于图像数据增强算法的处理方式会导致信息丢失，因此它们的性能不佳。

表 1　场景分类实验采用 ResNet-101 模型的测试结果比较

模型：ResNet-101	准确率(%)
无数据增强	59.60
Mixup	58.15
Cutmix	59.54
SpecAugment	57.68
SpecMix $\gamma = 0.3$	62.13

作者还就 SpecMix 算法中γ的取值进行了实验，如图 3 所示。结果表明，在 SpecMix 算法下，所有测试的γ取值都取得了优于其他方法的准确率，并且$\gamma = 0.3$时获得的准确率最高。

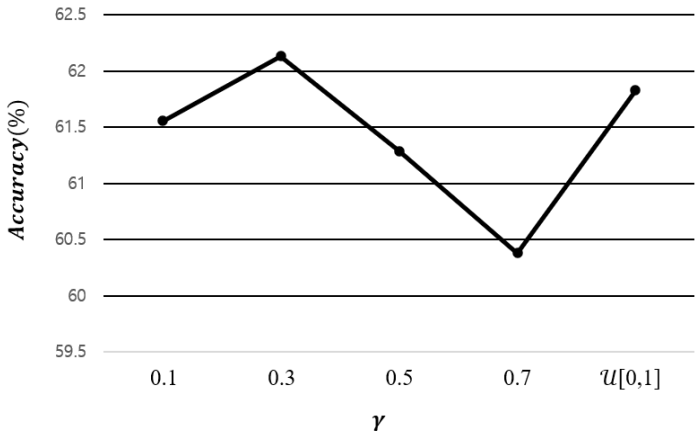

图3　场景分类实验中γ对SpecMix算法性能的影响

此外，采用 M. D. McDonnell 等人的模型进行测试的结果中，SpecMix 算法也取得了最好的结果，如表 2 所示（M 表示 M. D. McDonnell 等人的模型）。

表 2　场景分类实验采用 M. D. McDonnell 等人的模型的测试结果比较

模型：M	准确率(%)
无数据增强	68.90
Mixup	71.29
Cutmix	70.92
SpecAugment	69.07
SpecMix $\gamma = \mathcal{U}[0,1]$	71.60

作者利用此模型进行了 SpecMix 算法的消融实验，以探究掩蔽方式的影响。实验结果如表 3 所示。其中随机掩蔽指进行掩蔽时完全随机选取两组时频域特征图像的像素进行组合。结果表明，时间掩蔽和频率掩蔽都能增强 SpecMix 算法的性能。

表 3　掩蔽策略对 SpecMix 算法性能的影响

模型：M	准确率(%)
随机掩蔽	70.05
SpecMix(仅时间掩蔽)	70.42
SpecMix(仅频率掩蔽)	70.52
SpecMix	70.79

2.2 事件分类实验

在此实验中，作者采用 SECL UMONS 数据集进行评估。该数据集是一个室内录音数据集，用于声音事件的分类，包括拍手、敲门、键盘等 11 种事件。数据集包含 2178 个训练集片段和 484 个验证集片段。作者采用 ResNet-101 模型进行评估。SpecMix 算法的准确率达到 90.11%，同样高于其他数据增强算法，如表 4

中所示。

表 4　事件分类实验测试结果的比较

模型：ResNet-101	准确率(%)
无数据增强	96.07
Mixup	95.87
Cutmix	95.66
SpecAugment	96.90
SpecMix $\gamma = \mathcal{U}[0,1]$	97.11

2.3 语音增强实验

作者利用 Voicebank 数据集与多环境多通道噪声数据库（Diverse Environments Multichannel Acoustic Noise Database, DEMAND）进行了实验。Voicebank 与 DEMAND 分别提供了清晰的语音和有噪声的语音，采样率均为 48kHz。实验的评估指标包括语音质量的感知评估（PESQ）、信号失真的平均意见分数预测值（CSIG）、背景噪声干扰（CBAK）、总体信号质量（COVL）和分段信噪比（SSNR）。结果如表 5 所示。与其他数据增强策略相比，SpecMix 算法取得的效果最好。

表 5　语音增强实验测试结果的比较

	PESQ	CSIG	CBAK	COVL	SSNR
有噪声	1.97	3.35	2.44	2.63	1.67
无数据增强	2.50	3.44	3.18	2.95	9.26
Mixup	2.44	3.39	3.16	2.89	9.43
Cutmix	2.52	3.50	3.22	2.99	9.48
SpecAugment	2.41	3.48	3.16	2.93	9.35
SpecMix	2.54	3.60	3.24	3.05	9.57

作者测试了此实验中γ取值的影响，如图 5 所示。结果表明，无论γ取何种数值，使用了 SpecMix 算法的结果都优于无数据增强的结果。同样是在$\gamma = 0.3$时，SpecMix 算法取得的结果最好。

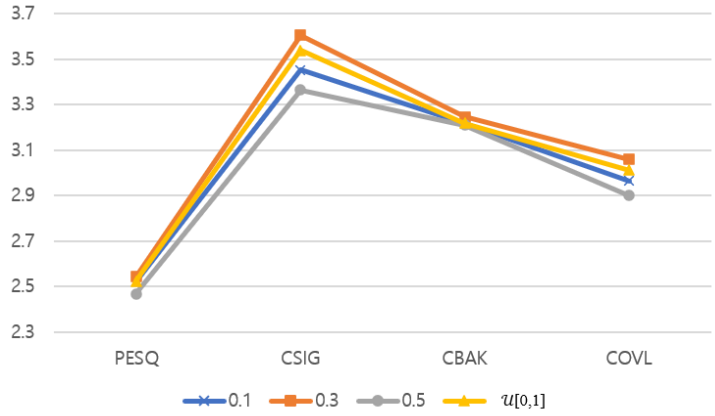

图5　语音增强实验中γ对SpecMix算法性能的影响

3. 结论

这篇论文提出了一种用于音频时频域特征数据增强的方法——SpecMix 算法。相比传统的时频域特征数据增强方法，此方法最大的优势是可以保留音频的频谱信息，避免数据增强带来的信息丢失，且计算过程较为简单。在多种模型与多种任务中进行的测试都表明使用 SpecMix 算法能够取得优于其他方法的结果。

五、启发思考

论文的实验部分给了我很大启发。首先，了解一个算法在不同实际应用中的效果是很有必要的。作者并没有只针对单一项音频处理任务来测试论文提出的算法，而是选取了声音场景分类、事件分类和语音增强多个具体任务进行测试。在声音场景分类中也采用了两种不同的深度学习模型进行实验。这些实验的目的就是充分测试算法的泛化能力。其次，在实验中除了整体算法以外，还要研究算法中不同组成部分各自的功能。作者对算法进行了消融实验，研究 SpecMix 算法中的两个关键部分：时间掩蔽和频率掩蔽的作用。结果证明，这两部分都能增强算法的性能。

深度学习的概念在这几年成为一股研究热潮，它在各种领域中得到了越来越广泛的应用。在进行深度学习的相关研究时，人们往往更关注针对某一具体任务进行深度学习网络模型的设计。从某种角度来说，这些工作确实发挥了重要作用，本文中就提到了计算机视觉领域中何凯明等人构造的 ResNet 模型。然而，很多追逐深度学习热潮的人没有意识到的是：深度学习想要获得更好的结果，必须有充足的数据量。没有一个足够大的数据集进行训练，得到的结果往往会出现过拟合。但是，一些实际问题中的处理任务由于各种葛塘的原因，不能获得足够的数据量。例如在事件检测中，如果检测目标是一些小概率事件，实际应用时就只能获得较少的数据量。这时，数据增强技术就显得尤为重要。实际上，这篇论文提出的数据增强方式并不复杂，计算过程也较为简单，但是却能获得很好的结果，这大概就是数据增强算法设计的魅力所在。今后，在进行与深度学习有关的研究时，我们不仅要关注模型设计，数据处理、数据增强这些所谓的“幕后”工作也同样重要。