arXiv:2305.02382v1 [cs.SD] 3 May 2023

使用预训练音频表征检测新颖且细粒度的声学序列

瓦苏达 *Kowtha*、*Miquel Espi Marques*、*Jonathan Huang*、*Yichi Zhang*、*Carlos Avendan* *o*

苹果

摘要

本研究探讨了用于少样本声音事件检测的预训练音频表征。我们重点研究新颖声学序列的少样本检测任务，即检测具有语义意义的时间结构的声音事件，且不假设能够获取非目标音频。我们开发了适用于预训练的表征方案，以及将这些表征迁移到少样本学习场景的方法。实验评估了预训练表征在 AudioSet 上的通用效用，并通过基于真实世界声学序列构建的任务，评估所提出的少样本方法的效用。我们的预训练嵌入适用于所提出的任务，并支持少样本框架的多个方面。

索引词— 小样本学习、声音事件检测、特征表示、声学序列

1. 引言

在声音事件检测（SED）任务中，小样本学习（FSL）范式为高效地利用少量样本，对用户单独指定的、具有独特特征的声音类别进行定义和检测提供了一条途径。这些声音类别不同于标准的带标签音频本体分类体系。在本研究中，我们探索了利用小样本学习检测声学序列的方法，例如音乐短句“pop-goes-the-weasel”。这类序列具有鲜明的时间结构：如果这一音乐短句在时间上受到严重遮蔽或被打乱，其类别身份就会发生变化（图1）。此前大多数关于小样本学习声音事件检测的研究主要关注粒度较粗的声音类别的检测或分类，在这种情况下，诸如“pop-goes-the-weasel”或“twinkle-twinkle-little-star”之类的特定序列可能都会被视为同一较粗粒度类别（如“铃声”或“音乐”）的成员。在这种典型的评估设置中，小样本学习方法主要根据其区分粗粒度目标声音事件与非目标声音类别的能力进行评估，而非目标声音类别通常与目标类别存在显著差异。相比之下，我们在具有挑战性的小样本细粒度评估设置中，评估了所提出方法检测特定序列的鲁棒性。在这一设置中，所有目标和非目标声音事件均属于同一个粗粒度类别。

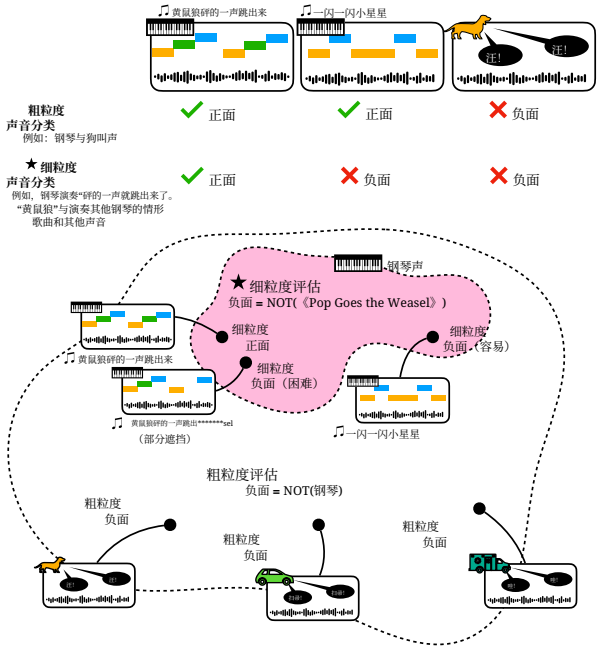

图 1. 声学序列与细粒度评估

我们通过联合设计合适的预训练骨干网络嵌入提取器的表示空间，以及利用这些表示的小样本学习方法的各个组成部分，来解决这一任务。所有这些都整合在一个统一的框架中（图 2），该框架具有以下特点：（1）无需使用专门或额外的数据来训练嵌入提取器；（2）小样本学习无需负面数据。我们认为，与通常针对通用弱标注声音事件表示而研究小样本学习系统的方法相比，这种方法具有创新性。

我们的贡献。我们展示了：（1）针对通用声音事件检测（SED）任务，有效预训练和蒸馏基于 CNN 的基线分类器的方法，并在 AudioSet 上取得了具有竞争力的基准性能；（2）实用的小样本学习-声音事件检测（FSL-SED）评估场景，在这些场景中，设计具有更高时间分辨率的预训练表示，其性能优于通过弱标签训练得到的声音事件检测嵌入，以及一种预训练此类表示的方法；（3）一种用于学习细粒度二元分类器的小样本学习框架，这些分类器通过弱标注音频进行预训练，无需假设能够获取非目标音频样本，并在具有挑战性的细粒度场景中进行评估，以检测特定的声学序列。

2. 相关工作

声音事件的小样本学习此前已有研究。例如，稀有声音 [1, 2],瞬态声音 [3], 人声 [4] 等特定场景已得到探索。本文提出了用于检测声学序列（尤其是更长序列，例如 >5 秒）的方法，而这一方向在相关文献中相对较少受到研究。

尽管近期图像领域的研究强调了利用适当的预训练表征空间的重要性 [5, 6, 7], ，但在小样本声音检测文献中尚未见到这一转变。这究竟是由于图像领域嵌入预训练的进展，还是由于元学习方法对分布偏移的某些方面较为脆弱 [5]

尚不清楚。已有研究表明，音频特有的因素（如多声部性、声学退化和背景噪声）以及弱标注数据，可能导致音频和视觉领域都缺乏一种适用于所有情况的方法 [8]。然而，先前的研究尚未探讨将表征空间与小样本算法组件协同设计的方法，也未充分处理非目标音频（负面）不可用等问题。因此，我们研究了精心选择的预训练表征的潜力，在评估其对领域偏移的鲁棒性的同时，采用务实的方式处理混响、弱标注目标以及多声部性的相关问题等音频特有因素。

3. 音频表征预训练

我们基于标准的预训练多类声音事件分类器，构建一种灵活的表征，使其能够区分细粒度的声学序列。

训练数据。已有多个值得关注的数据集专为声音分类而开发 [9], [10], [11], [12]。对于小样本学习，理想的表征应涵盖种类丰富的声音；而要训练出稳健的表征，大量数据至关重要。我们选用了包含 527 类声音的 AudioSet [11], , 其中约有 200 万个训练样本和 2 万个评估样本。包括 AudioSet 在内的典型声音分类数据集都是**弱标签**数据集：只要某一类别的声音事件在文件中的任意位置出现，就会为该文件标注相应标签。相比之下，使用帧级目标会更理想，这样训练出的模型便能表征音频中的时间动态。尽管可以获取**强标签**

尽管可以通过标注每个声音事件的起始和结束时间来实现这一目标，但对于 AudioSet 这种规模的数据集而言，强标注的成本高得令人难以承受。

3.1. 模型架构

我们训练了三种卷积神经网络架构：1）一个参考模型，其训练目标是最大限度地提升在弱标注 AudioSet 训练目标上的性能；2）一个更小的 MobileNetV2 模型，其尺寸适合在移动设备上运行，并通过较大的教师模型进行蒸馏

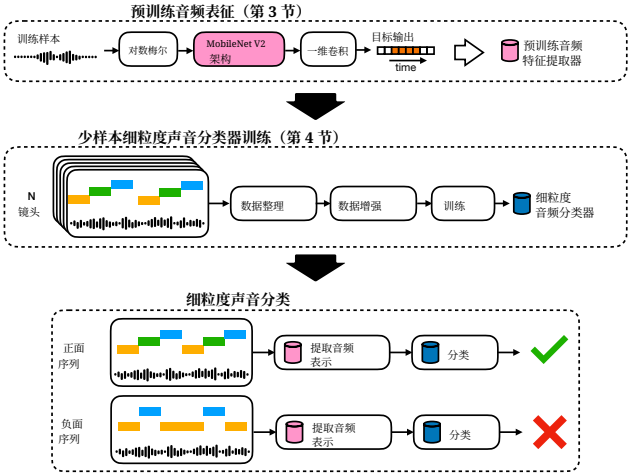

Рис. 2. Обзор процесса обучения и вывода модели

，同时也使用弱标注目标进行训练；3）一个模型，其卷积神经网络架构与 2）相同，但采用适合生成强标签预测估计值的输出层。所有这些架构都支持推理时灵活调整输入时长。卷积神经网络的输入为 64 维 logmel 频谱图，由 25 毫秒输入帧、10 毫秒帧移提取而来。

参考模型。我们使用 ResNet-50 [13]骨干网络来最大化声音分类性能。ResNet 的输出按通道进行池化，从而将任意长度的输入音频转换为 2048 维输出向量。池化后的向量经过两个全连接层，其中第一个层使用 ReLU 激活函数，最终输出 527 个类别。池化层将所有时间步的信息聚合为一个固定维度的向量，并为整个 10 秒训练文件生成单一预测输出。虽然池化层在弱标签训练目标下能够取得良好结果，但我们发现，以这种方式训练的嵌入在时间序列的小样本学习方面存在局限性。

知识蒸馏嵌入。对于设备端推理，理想情况下应使用在性能与复杂度之间具有良好权衡的模型。我们对 MobileNetV2[14] 进行了修改，将最后一个卷积层的通道维度从 1280 调整为 512，并将其命名为 MobileNetV2-S。随后，我们采用知识蒸馏（KD） [15] 进一步提升其性能。我们使用一个更大的 MobileNetV2-L 作为教师模型，该模型在 MobileNetV2-S 的每一层上应用了 1.4 倍的宽度乘数，并将最后一个卷积层的通道数设为 1,500。学生模型 MobileNetV2-S-KD 通过同时最小化教师模型与学生模型 logits 之间的 K-L 散度以及学生模型的类别交叉熵损失进行训练。尽管其规模仅为参考模型的 $\sim$1/13，但仍取得了良好的性能（见第 5 节）。检测器的设计与上述模型类似，不同之处在于嵌入维度有所不同，并且全连接层的数量减少至 1 层。

强标签嵌入。虽然池化层在

前两种架构提供了一种利用弱标注数据集进行训练的方法，但其潜在缺点是无法保留音频中的时间动态。如果存在强标注数据集，我们就可以创建一个能够进行帧级预测的输出层。在强标签嵌入中，我们保留了相同的 MobileNetV2-S 主干网络，但仅沿频率维度进行池化，而不沿时间维度进行池化，从而保留时间维度。池化后的输出被输入一个 1x1 一维卷积，该卷积每隔 320ms 生成 527 个类别的输出。该模型被命名为 MobileNetV2-S-Strong，其训练基于 MobileNetV2-S-KD 初始化。

3.2. 模型训练

弱标签训练。我们采用了先前工作中的各种已知音频分类训练技术 [16], [17]: 1) 在 +/-10%范围内对音频进行随机重采样，2) SpecAugment [18], 3) mixup [19], 4) +/-20 dB随机增益，以及5) 在数据加载器中使用类别不平衡采样，以增加少数类的出现次数。我们使用AdamW优化器 [20]，采用单周期学习率调度，将学习率预热至0.01，然后在最后一个周期30衰减至0.0001。通过这些设置，我们为 ResNet 教师模型取得了令人满意的结果。弱标签嵌入使用了几乎相同的流程进行训练，但额外使用了一个来自教师模型输出的训练目标。

伪强标签训练。帧级训练目标用于训练强标签嵌入。然而，执行强标签标注是一项劳动密集、耗时且成本高昂的任务。因此，我们转而利用弱标签参考模型生成伪强标签。通过实验，我们发现，当输入时长远短于 10 秒的训练样本时，使用弱标签训练的模型仍能产生相当不错的预测结果。对于每个 10 秒的 AudioSet 文件，我们以 100 毫秒为步长生成 ResNet 模型的预测结果，输入时长为 0.5 秒。我们采用 0.5 的阈值，将每个类别的输出转换为二值结果。随后，使用这些伪标签训练强标签嵌入。除训练目标外，训练流程与弱标签训练相同。

4. 应用：少样本检测

音频整理。我们不假设少样本声音事件已经分段，而是利用嵌入来发现目标声音及其边界。这使我们无需起始点或结束点标注，也能从音频中进行少样本学习。

给定 K 段任意长度的音频，每段都包含一个目标声音实例，我们需要估计一组起始和结束时间戳 $D_{enroll} = \{(\mathbf{x}_i,\ y_i)\}_{i=1}^{K}$，且不具备关于目标的任何先验信息。我们采用了以下方法：

1. 在每个音频镜头中找到响亮片段。使用以 $loud$ 标签（能量最高的前 5% 对数梅尔帧）和 $quiet$训练的逻辑回归模型（最低的 5%）用于在每个镜头中分别找出被分类为 $loud$的片段。

2. 找出在声音表征空间中语义相似且出现在多个片段中的片段。利用成对固定长度嵌入之间的余弦距离对片段进行分组，并且每段只保留最早的起始点和最晚的结束点，舍弃其余部分。

3. 以最短片段作为范例，通过在 logmel 域中寻找与范例等长且互相关最高的子片段，将其余片段调整为相同长度。

二分类器训练方法。给定强标签数据集 D_{enroll}，我们的目标是生成增强数据集$D_{train} = \{(e(\mathbf{x}_i)\ y_i)\}_{i=1}^{K+A}$，其中 $K >> A$。这需要通过增强正样本并生成负样本来提高类内方差，以近似现有目标样本集合中缺失的、已知会出现在推理时条件中的某些方面。随后，我们使用增强后的$D_{train} = \{(e(\mathbf{x}_i)\ y_i)\}_{i=1}^{K+A}$ 来训练一种计算开销低且具有良好泛化能力的二分类器。为此，我们首先使用预训练嵌入 $e(\cdot)$，从分段音频中提取预训练的可变长度嵌入序列作为特征。

目标类别嵌入增强。我们希望增强目标类别。我们在时域中进行处理：将音频整理的分段输出（起始点、结束点）略微扩大 500 毫秒，并在扩大的边界内进行时间平移。我们还在嵌入域中进行增强，将 Δ-编码器 [21] 适配到时域：把 1024 维嵌入序列视为固定长度嵌入的集合。形变通过嵌入帧对进行学习。我们假设嵌入对来自成对的˜声音事件，其形式为 X、 X。因此，学习到的形变 z_i可解释为嵌入域中的一种操作：该操作根据成对的干净样本$e_T X$合成声学退化的 $e_T X$；它是嵌入域中将干净声音事件样本与房间脉冲响应（RIR）进行卷积的对应操作。

非目标类别嵌入合成。为获取负面示例，我们对增强后的目标序列应用参数化的嵌入域掩蔽和打乱操作。通过屏蔽适当长度的连续嵌入帧，或打乱适当长度的块（两者均取决于目标序列的长度），我们以最小程度扰动目标音频表示。因此，合成出的负面示例将接近最终决策边界。

学习。最后，我们提出一种训练小型模型的方法，使其学习具有泛化能力的决策边界。我们在单个端到端模型中实现了可学习的线性投影以降低维度，利用一维膨胀因果卷积进行时序建模，并采用二元分类器；训练时使用多目标损失，要求输入和中间特征图具有大间隔损失相对于。我们所采用的决策边界距离公式（1）是对先前提出的一种方法的改进

大间隔损失 [22]。结合一项较小的二元交叉熵损失，这种多目标损失有助于提升模型在较长序列上的学习效果。

$$\tilde{d_n} = \frac{f_i(x_n) - f_{\neg i}(x_n)}{\|h_T[\nabla_{x_n} f_i(x_n) - \nabla_{x_n} f_{\neg i}(x_n)]\|_2^F + \epsilon}, \quad (1)$$

5. 实验

我们采用 AudioSet 和内部收集的自然家庭声音事件与场景数据集进行评估。具体的声学序列取自后一个数据集，这是由于在典型的带标注声音分类语料库中，具有独特时间特性的声音事件类别缺乏相关的更细粒度子类标签。所有数据集均重采样至 16 千赫。

表1。声学表征预训练方法：域内评估（AudioSet）

系统	参数	mAP	d-prime
参考文献	25.6M	0.439	2.72
MobileNetV2-L	4.45M	0.394	2.63
MobileNetV2-S	1.98M	0.376	2.53
MobileNetV2-S-KD	1.98M	0.401	2.67
MobileNetV2-S-Strong	1.98M	0.254	2.43

预训练模型基准测试。在表1中，我们报告了各种嵌入的声音分类结果。首先，参考模型的平均精度均值为0.439，优于14层PANN [17]。其次，基于蒸馏的MobileNetV2-S-KD比参考模型小得多；与MobileNetV2-L教师网络相比，模型压缩率达2.2倍。令人惊讶的是，学生模型（平均精度均值为0.401）优于教师模型（平均精度均值为0.394）。第三，未经蒸馏的MobileNetV2-S取得了0.376的平均精度均值。这表明从教师模型中提取信息十分重要。最后，我们展示了具有精细时间分辨率强标签嵌入的MobileNetV2-S-Strong的平均精度均值和d-prime。请注意，该模型使用由参考模型预测结果生成的100毫秒帧级伪强标签。以这种方式估算的标签可能与弱标签真实值存在偏差。然而，由于强标签嵌入能够学习到理想的精细时间细节，我们仍将MobileNetV2-S-Strong用于少样本检测，下一小节的结果将展示其优势。

所提框架分析。我们在一项分布外（OOD）检测任务中评估细粒度鲁棒性，该任务涉及两个实际假设：(a) 正样本的类内方差无法匹配推理时条件；(b) 不假设能够访问非目标样本。

每项推理时任务由三个目标样本和一百八十个非目标样本组成，对应于一个目标序列与九个未见

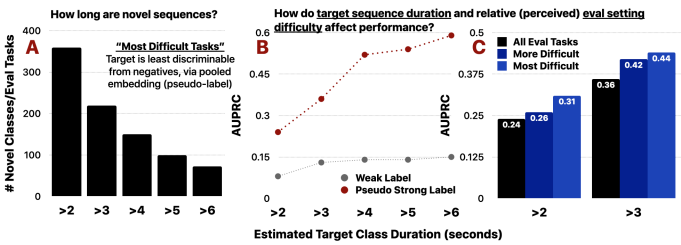

图3。（A）所评估新颖序列的时长分布（通过整理数据估算）。（B）时长较长的目标序列可获得更大的相对性能提升（指标：每个注册序列重复10次所得平均AUPRC的中位数）。（C）目标序列长度与评估设置难度。

非目标序列。评估集中每个序列的录音中，近场和远场录音各占 50%。非目标样本的声学内容因评估任务而异，与录音所在的家庭环境相对应。

图3(A)展示了所评估目标序列的数据分布。图3(B)展示了所提少样本方法的性能，其中伪强标签嵌入（PSL）总体优于相应的弱标签嵌入（WL）：随着目标序列时长增加，我们观察到基于PSL的少样本系统带来了更大的相对提升。这表明，对于迄今文献中探索极少的长时长声学序列检测任务，所提预训练嵌入与所提系统相结合具有优势。最后，我们通过“任务难度指数”衡量不同任务的相对难度：该标量难度度量根据定长嵌入之间的余弦距离计算得出。如图3(C)所示，评估任务越困难，所提系统带来的相对收益越大。这表明我们提出的方法与更广泛研究的定长嵌入小样本学习方法可能具有互补性。未来，我们计划研究这类方法的互补性。

6. 结论

在本研究中，我们提出了一种用于新颖声学序列少样本检测的统一方法，适用于资源受限的使用场景。本研究包含声学表征预训练方法，以及在此基础上设计的小样本学习系统组件。我们提出的嵌入预训练、蒸馏和伪强标签标注基准优于现有方法。我们使用预训练音频表征对声学序列进行小样本学习，并发现所提出的表征能够支持针对尚未充分探索的任务开展少样本声音事件检测，同时在具有挑战性的条件下进行了评估。

7. 参考文献

[1] K. Shimada、Y. Koyama 和 A. Inoue，《利用背景噪声类别进行稀有声音事件少样本检测的度量学习》，*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），2020 年。

[2] C. Zhao、J. Wang、L. Li、X. Qu 和 J. Xiao，《用于检测罕见声音事件的自适应小样本学习算法》，*arXiv*，2022。

[3] S. Chou、K. Cheng、J. R. Jang 和 Y. Yang，《利用注意力相似性学习匹配瞬态声音事件，以实现小样本声音识别》，*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2019。

[4] Y. Wang、J. Salamon、N. J. Bryan 和 J. P. Bello，《小样本声音事件检测》，*IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2020。

[5] W. Chen、Y. Liu、Z. Kira、Y. F. Wang 和 J. Huang，《进一步审视小样本分类》，国际学习表征会议（*ICLR*），2019。

[6] Y. Tian、Y. Wang、D. Krishnan、J. B. Tenenbaum 和 P. Isola，“ 重新思考少样本图像分类：一个好的嵌入就是你所需要的一切吗？ ”，*ECCV*，2020。

[7] C. Liu、C. Xu、Y. Wang、L. Zhang 和 Y. Fu，“ 一种出奇简单的单样本学习基线”，*IEEE/CVF*计算机视觉与模式识别会议研讨会（*CVPRW*），2020。

[8] Y. Wang、N. J. Bryan、J. Salamon、M. Cartwright 和 J. P. Bello，“谁说了算？重新思考音频的少样本学习”，*IEEE*信号处理在音频与声学中的应用研讨会（*WASPAA*），2021。

[9] Karol J Piczak，“ESC：用于环境声音分类的数据集”，载于第*23*届*ACM*国际多媒体会议论文集，2015年，第1015–1018页。

[10] Honglie Chen、Weidi Xie、Andrea Vedaldi 和 An- drew Zisserman，“Vggsound：大规模视听数据集”，载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第721–725页。

[11] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter，“ Audio set：音频事件本体与人工标注数据集”，载于*2017*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2017年，第776–780页。

[12] Eduardo Fonseca、Xavier Favory、Jordi Pons、Frederic Font 和 Xavier Serra，“Fsd50k：人工标注声音事件的开放数据集”，发表于*IEEE/ACM*音频、语音与语言处理汇刊，第30卷，第829–852页，2021年。

[13] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun，“用于图像识别的深度残差学习”，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第770–778页。[14] Mark Sandler、Andrew Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen，“MobileNetV2：反向残差与线性瓶颈”，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第4510–4520页。

[15] Geoffrey Hinton、Oriol Vinyals 和JeffDean，“神经网络中的知识蒸馏”，*arXiv*预印本 *arXiv:1503.02531*，2015年。

[16]Jonathan J Huang 和 Juan Jose Alvarado Leanos，“ Aclnet：高效的端到端音频分类卷积神经网络”，*arXiv*预印本 *arXiv:1811.06669*，2018 年。

[17] Qiuqiang Kong、Yin Cao、Turab Iqbal、Yuxuan Wang、Wenwu Wang 和 Mark D Plumbley，“Panns：用于音频模式识别的大规模预训练音频神经网络”，*IEEE/ACM*音频、语音与语言处理汇刊，第28卷，第2880–2894页，2020年。

[18] Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le，“Specaugment：一种用于自动语音识别的简单数据增强方法”，*arXiv* 预印本 *arXiv:1904.08779*，2019 年。

[19] Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz，《mixup：超越经验风险最小化》，*arXiv* 预印本 *arXiv:1710.09412*，2017 年。

[20] Ilya Loshchilov 和 Frank Hutter，《修正 Adam 中的权重衰减正则化》，*ICLR 2018*年，2018年。

[21] Eli Schwartz、Leonid Karlinsky、Joseph Shtok、Sivan Harary、Mattias Marder、Abhishek Kumar、Rogerio Feris、Raja Giryes 和 Alex Bronstein，《 Delta-encoder：一种用于少样本目标识别的有效样本合成方法》，第*32*届神经信息处理系统会议（*NeurIPS*），2018年。

[22] Gamaleldin Elsayed、Dilip Krishnan、Hossein Mobahi、Kevin Regan 和 Samy Bengio，《用于分类的大间隔深度网络》，第*32*届神经信息处理系统会议（*NeurIPS*），2018年。