哈爾濱工業大學

课程报告

题　　目 SELF-SUPERVISED SPEAKER RECOGNITION WITH LOSS-GATED LEARNING

专　　业 人工智能

学　　号 1190202107

班　　级 1903602

姓　　名 姚舜宇

报 告 日 期 2022.10.4

论文标题

Self-supervised speaker recognition with loss-gated learning

作者单位：NUS，A*STAR，Fortemedia Singapore，UEF，CUHK-ShenZhen

摘要

在自监督学习的说话人识别中，伪标签作为监督信号有很大的作用。但由于伪标签的不可靠性，用于说话人识别的模型往往无法总是从伪标签中学习到有用的信息。在本工作中，作者观察到说话人识别网络对于可靠标签数据的建模比不可靠标签要更快速。这个猜想促使作者提出一个门限学习（LGL）机制，用于在网络的训练中依据模型对数据的拟合能力来确定标签的可靠性。通过 LGL 机制，作者提出的说话人识别模型获得了 46.3%的性能提升。更多的实验表明，该方法在 VoxCeleb1 和 VoxCeleb2 数据集取得了很好的性能。

关键词：自监督说话人识别，伪标签选择，门限学习

介绍

说话人识别的目标是从说话声音来识别人。近年以来，说话人识别模型往往通过监督学习来训练，并且取得了很好的性能。然而，这些方法通常需要大量人工标注的数据来训练，这些数据的代价非常昂贵。自监督学习不需要数据标签既可以训练，可以节省大量的标注成本，利用海量的无标签数据。

目前的自监督说话人识别的 SOTA 方法由两个阶段组成。在阶段一，我们使用对比学习的方法如 SimCLR，MoCo，来让说话人编码器学习到有意义的语音表示。与此同时，一些损失函数会用于设定对比标签并且优化编码器。然而，由于缺乏说话人身份信息，阶段一的性能非常有限。近期，可以迭代的阶段二用于缓解这个问题。在这里，对于每一个话语段，使用聚类算法来生成伪标签。有了伪标签，网络就可以用监督的模式来训练一个分类器。为了提升说话人编码器的效果，这个过程可以重复多次。

目前的 SOTA 方法把伪标签用于全监督的分类，因此，阶段二中分类的质量会决定自监督说话人识别的性能上限。通常，伪标签中包含很多不可靠的信息，这些错误标签会影响编码器的性能，这也强调了有效并可靠地选取伪标签的重要性，这也与标签平滑类似。

在本工作中，作者假设神经网络对有可靠标签的数据建模要比那些标签不可靠的数据要更快。具体的，考虑某一时刻的一个话语段，当它和伪标签的损失较小时，可以认为伪标签是可靠的，否则不可靠。作者设计了一个小实验来验证这个假设。为此，作者提出了文献学习 LGL 策略来选择有可靠伪标签的数据，具体来说，用一个阈值来区分损失，表示数据的可靠性，只有可靠的数据才能用于更新网络。总的来说，贡献可以总结为：

- 在自监督学习中，作者确认了神经网络对有可靠标签的数据比不可靠标签的数拟合地更快。
- 基于此发现，作者提出门限学习 LGL 策略来有效选取有可靠标签的数据。

基线工作

在这里作者描述了基线的两阶段结构。如下图所示，说话人编码器用对比学习来训练，然后通过聚类并且先努力一个分类网络来得到伪标签。

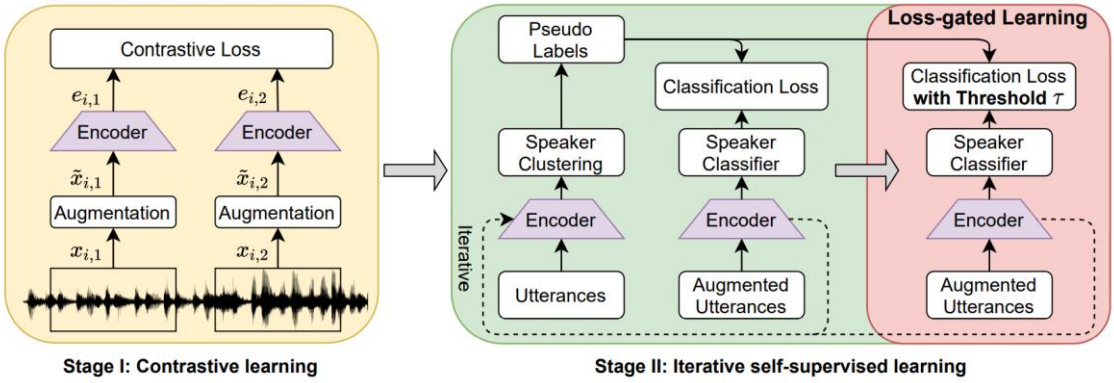


Fig. 1. Framework of self-supervised speaker recognition with loss-gated learning.

阶段一：对比学习

在阶段一中，作者根据对比学习设计了一个自监督的预训练任务。在每个批次中，随机选取 N 个无标签的话语段$x_1, x_2, \cdots, x_N$，如上图，随机考虑话语段x_i的两个不重合的相同长度的子语段$x_{i,1}, x_{i,2}$。然后使用随机噪声增强得到$\tilde{x}_{i,1}, \tilde{x}_{i,2}$。然后送入说话人编码器$f(\cdot)$来得到语音编码$e_{i,j} = f(\tilde{x}_{i,j})$，这里$i \in \{1, \cdots, N\}$，$j \in \{1, 2\}$。这里假设每个话语段里面只有一个说话人，所以从一个话语段里面分出的两个子段可以视为正样本对。另外，从其他话语段中分出的子段可以作为一个负样本。为了让正样本对之间更加接近，负样本对之间更远离，可以针对每个正样本对和负样本对定义对比损失函数：

$$l_{i,j} = -\log \frac{\exp(\cos(e_{i,1}, e_{i,2}))}{\sum_{k=1}^{N} \sum_{l=1}^{2} \mathbb{1}_{\substack{k \neq i \\ j \neq l}} \exp(\cos(e_{i,j}, e_{k,l}))}$$

放到每个批次中，损失函数则表示为：

$$L_{\text{scl}} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2} l_{i,j}$$

上式中的$\cos(\cdot,\cdot)$表示余弦相似度。通过最小化该损失函数，编码器可以学习到区分正负样本的话语段表示。

阶段二：迭代自监督学习

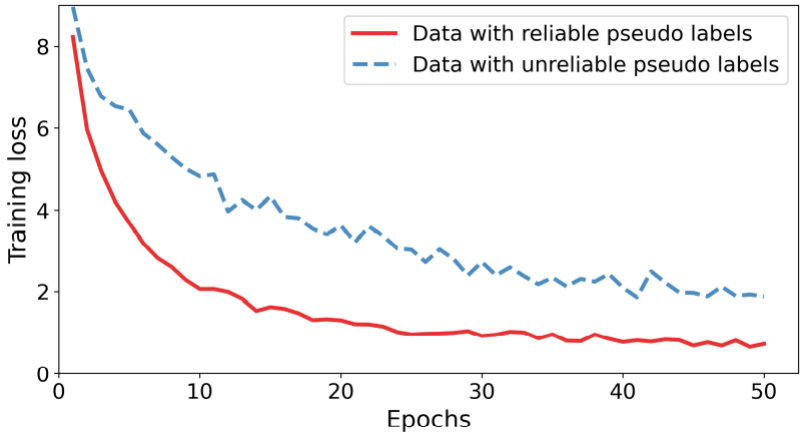
阶段二如上图所示。首先，使用阶段一训练出的编码器作为初始模型，来对每个话语段提取说话人编码。基于这些编码，作者使用 kmeans 聚类来生成伪标签，每个类别的话语段认为是属于同一个说话人。然后，再用这些伪标签来重新训练编码器。用伪标签和编码器的分类做损失，选取可靠度高的伪标签重新训练，重复这些步骤直到系统收敛。注意到每轮迭代中用于说话人分类的编码器会在下一轮迭代中生成用于聚类的语音编码。

门限学习

首先，需要定义阶段二中的伪标签的可靠性。我们可以使用匈牙利算法获得一个聚类伪标签和正确标签的一一映射。如果正确标签和伪标签相同，则定义该数据可靠，否则不可靠。但是在自监督学习中，无法提供这样准确的信息。

传统的，基线系统使用伪标签来训练阶段二，没有考虑到伪标签的可靠性，因此丢失了一定的性能。这促使作者考虑使用具有可靠标签的数据来训练一个更优的编码器。

问题在于如何有效选取可靠的伪标签。作者做了简单的实验，数据表明，模型对于可靠数据和不可靠数据的学习能力和收敛速度不同，如下图所示。



从上图中可以得出，当数据的标签可靠时，训练损失会下降地更快，验证了假设。

基于此，作者提出了门限学习 LGL 策略。在阶段二获得伪标签后，相比于传统方法中将所有数据送入下一轮迭代的训练中，作者筛选出可靠的数据来继续训

练。具体的，作者引入了一个阈值τ：

$$L_{spk} = \sum_{i=1}^{N} l_i \mathbb{1}_{l_i<\tau}$$

这里的l_i表示一个数据点的训练损失。假设是，在一些轮数的训练后，具有小损失值的数据会比其他具有大损失值的数据更可靠。因此，LGL 只用了那些小损失的数据来训练网络。作者同时训练编码器和分类器，直到系统达到最佳性能。然后使用这个编码器进入下一个迭代中进行聚类。总之，作者提出的方法和传统方法都是有两个阶段，LGL 策略用于在阶段二中解决不可靠伪标签的问题。

实验

本段主要描述实验设置与超参数。作者使用 VoxCeleb2 作为训练集，VoxCeleb1 作为测试集，不使用任何的标签。训练中，批次大小设置为 256，使用 Adam 优化算法，使用适当数据增强。阶段一中，初始学习率设置为 0.001，并且每 5 轮下降 5%。阶段二中，使用 kmeans 算法来聚类，根据前人的工作，类别数设置为 6000。

结果与分析

和现有工作的比较

下表描述了作者提出的方法和现有工作的性能比较。使用 LGL 的自监督说话人识别系统在 Vox_O 中取得了 1.66%的错误率，比现有的最好方法高出 20.95%，同时在 Vox_E 和 Vox_H 数据集上也有很大提升。

Stage	Method	Vox_O	Vox_E	Vox_H
I	Nagrani et al. [12]	22.09	-	-
	Chung et al. [13]	17.52	-	-
	Inoue et al. [35]	15.26	-	-
	Huh et al. [21]	8.65	-	-
	Zhang et al. [20]	8.28	-	-
	Xia et al. [18]	8.23	-	-
	Mun et al. [19]	8.01	-	-
	Ours	7.36	7.90	12.32
II	Cai et al. [14]	3.45	4.02	6.57
	Ours w/o LGL	3.09	3.81	6.32
	Thienpondt et al. [15]	2.10	-	-
	Ours with LGL	**1.66**	**2.18**	**3.76**

LGL 的作用

作者总结了每轮迭代是否使用 LGL 的性能差异。可以发现 LGL 可以迅速提升系统性能，并且在前期提升更为明显，因为迭代次数较少时，不可靠数据占比会更多。

Iteration-#	1	2	3	4	5
W/o LGL	4.92	4.00	3.68	3.22	3.09
With LGL	3.52	2.41	2.07	1.95	**1.66**

消融实验

作者在本节通过实验验证了聚类数量和阈值选取的差异。

# Clusters	3,000	4,500	6,000	7,500	9,000
W/o LGL	5.29	4.89	4.92	5.05	5.31
With LGL	3.49	3.35	3.52	3.48	3.71

Threshold (τ)	1	2	3	4	5	$+\infty$
EER	**3.52**	3.77	3.74	4.10	4.14	4.92

另外，作者通过衡量聚类结果的好坏，证明了 LGL 策略的有效性。如下图，NMI 表示归一化互信息，越高表示聚类结果越好。(a)图表示，随着迭代轮数的增加，使用 LGL 策略后，对数据的聚类结果远优于不适用 LGL。(b)图表示，使用 LGL 策略筛选出的数据的聚类结果一直好于全部数据的聚类结果，这也说明了 LGL 能够选择出可靠的数据来帮助网络的训练。