计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(2): 336−345, 2025

面向合成语音检测的弹性正交权重修正连续学习算法

张晓辉[1,2] 易江燕[1] 陶建华[3] 周俊佐[1]
[1](中国科学院自动化研究所 北京 100190)
[2](北京交通大学计算机与信息技术学院 北京 100044)
[3](清华大学自动化系 北京 100084)
(21120320@bjtu.edu.cn)

Elastic Orthogonal Weight Modification Continual Learning Algorithm in the Context of Synthetic Speech Detection

Zhang Xiaohui[1,2], Yi Jiangyan[1], Tao Jianhua[3], and Zhou Junzuo[1]
[1] (*Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2] (*School of Computer and Information Technology*, *Beijing Jiaotong University*, *Beijing* 100044)
[3] (*Department of Automation*, *Tsinghua University*, *Beijing* 100084)

Abstract Currently, deep learning has achieved significant success in the field of synthetic speech detection. However, deep models commonly attain high accuracy on test sets that closely match their training distribution but exhibit a substantial drop in accuracy in cross-dataset scenarios. To enhance the generalization capability of models on new datasets, they are often fine-tuned with new data, but this leads to catastrophic forgetting, where the model's knowledge learned from old data is impaired, resulting in deteriorated performance on the old data. Continuous learning is a prevalent approach to mitigate catastrophic forgetting. In this paper, we propose a continuous learning algorithm called elastic orthogonal weight modification (EOWM) to address catastrophic forgetting for synthetic speech detection. EOWM mitigates knowledge degradation by adjusting the direction and magnitude of parameter updates when the model learns new knowledge. Specifically, it enforces the updates' direction to be orthogonal to the data distribution of the old tasks while constraining the magnitude of updates for important parameters in the old tasks. Our proposed algorithm demonstrates promising results in cross-dataset experiments within the domain of synthetic speech detection. Compared with fine-tuning, EOWM reduces the equal error rate (EER) on the old dataset from 7.334% to 0.821%, representing a relative improvement of 90%, and on the new dataset, it decreases EER from 0.513% to 0.315%, corresponding to a relative improvement of 40%.

Key words synthetic speech detection; continual learning; elastic orthogonal weight modification; pre-trained model; deep neural network

摘 要 目前,深度学习在合成语音检测领域取得了巨大的成功.然而,通常情况下,深度模型可以在与训练集分布相似的测试集上取得高准确率,但在跨数据集的情境下,其准确率却会显著下降.为了提高模型在新数据集上的泛化能力,通常需要对其进行微调,但这会导致模型遭受灾难性遗忘.灾难性遗忘指的是模型在新数据上的训练会损害其从旧数据中获得的知识,导致对旧数据的识别性能下降.目前,克服灾难性遗忘的主要算法之一是连续学习.面向合成语音检测提出了一种连续学习算法——弹性正交权重修正

收稿日期: 2023-04-10; **修回日期**: 2024-01-08
基金项目: 国家重点研发计划项目(2020AAA0140003); 国家自然科学基金项目(61831022, U21B2010, 61901473, 62006223, 2101553)
This work was supported by the National Key Research and Development Program of China (2020AAA0140003) and the National Natural Science Foundation of China (61831022, U21B2010, 61901473, 62006223, 2101553).
通信作者: 易江燕(jiangyan.yi@nlpr.ia.ac.cn)

（elastic orthogonal weight modification, EOWM），用于克服灾难性遗忘. 该算法通过修正模型在学习新知识时的参数更新方向和更新幅度，以减少对已学知识的损害. 具体来说，该算法在模型学习新知识时要求参数的更新方向与旧任务的数据分布正交，并同时限制对旧任务中重要参数的更新幅度. 在合成语音检测领域的跨数据集实验中，算法取得了良好的效果. 与微调相比，该算法在旧数据集上的等错误率（equal error rate, EER）从 7.334% 降低至 0.821%，相对下降了 90%；在新数据集上的等错误率从 0.513% 降低至 0.315%，相对下降了 40%.

关键词 合成语音检测；连续学习；弹性正交权重修正；预训练模型；深度神经网络

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202330311 **CSTR:** 32373.14.issn1000-1239.202330311

随着语音合成和语音转换技术的发展，合成语音已经能与真实语音相媲美，这使得人耳很难将二者区分开. 虽然该技术给人类生活带来极大的便利，但是不良用途的合成语音给国家和社会带来重大安全危害，因此，合成语音检测成为了近年来的一个重要的研究方向. 2013 年，语音国际会议 INTERSPEECH 的 Special Session 中正式提出了针对说话人认证系统的生成语音反欺诈问题，旨在从混有真实语音和合成语音的数据集中，通过机器学习等技术，将二者区别开，并将该技术用于解决合成语音攻击问题. 为了应对合成语音攻击，学术界和工业界近年来举办了多场相关比赛，旨在提高模型的合成语音检测能力. 自 2015 年开始到 2021 年，INTERSPEECH 已经连续举办了 4 届针对说话人认证系统语音欺骗攻击挑战赛（ASVspoof automatic speaker verification spoofing and countermeasures challenge, ASVspoof）[1-4]. 2022 年，中国科学院自动化研究所联合新加坡国立大学在语音国际会议 ICASSP 上举办语音深度合成检测挑战赛（audio deep synthesis detection challenge, ADD）[5]. 其中，在 ASVspoof 比赛中脱颖而出的模型往往都在当次比赛测试集上具有很高的准确率. 然而，在跨数据集场景下，即用旧数据集的训练集训练的模型在新数据集的测试集上测试，这些检测模型的识别准确率往往会有明显的下滑[6-7]. 这主要是因为不同数据集的特征分布不匹配[6-7]，导致在旧数据集上训练的模型在新数据集上识别能力降低.

解决跨数据集分布不匹配的问题最常见的算法，是对在旧数据集上训练完毕的模型用新数据集的训练数据集微调，以提高在新数据集上的识别准确率. 然而，微调也会带来灾难性遗忘的问题，即微调后的模型在旧数据集上的识别准确率会有较大的降低. 目前已经有多种连续学习算法在多个领域的克服遗忘方面取得了一定的成果[8-9]. 在合成语音检测领域中，存在一部分针对该领域设计的克服灾难性遗忘的算法，以及部分直接应用于该领域的经典的连续学习算法. 其中经典的连续学习算法大多在其他领域，例如计算机视觉、目标检测、特征选择等取得了一定的成果[10-15].

1）经典的连续学习算法

该算法依据的是不同算法的核心思想，连续学习的算法分为 3 类：动态架构（dynamic architectural, DA）算法、基于经验重放（memory-based, MB）的算法和正则化（regularization）算法. 动态架构算法通过在新任务的学习中引入动态变化的模型结构，如附加模型层、节点或模块等，来存储新获取的知识，旧知识则被保存在旧的网络模块中，不会因为新的学习过程被损害. 然而，这种算法的模型容量会随着连续学习过程中任务数的不断增加而显著提升，通常是指数级增长，这对系统的内存和存储提出了挑战. 渐进网络（progressive network）[16-17] 和动态扩展网络（dynamically expandable network, DEN）[18] 作为代表性的动态架构算法，通过复制基础网络中的部分节点，并将新节点与旧网络横向链接来学习新知识. 基于经验重放的算法通过保存一部分旧数据，在新任务的学习中加入旧数据，从而使模型在获取新知识的同时尽量减少对旧知识的遗忘. 梯度情景记忆（gradient episodic memory, GEM）算法[19] 保存旧任务中最后出现的一部分数据，并引入到新任务的训练中. 这种算法通常需要在新任务的训练过程中额外引入蒸馏损失（distillation loss）或其他度量 2 个深层表示之间相似性的算法[20-22]，并且需要根据学习任务优化保存的旧任务的数据类型. 正则化算法通常在新任务的学习中对新函数引入正则化或直接对模型参数进行修正，其优点是不需要保存旧的数据，没有额外引入新的节点. 正交权重修正（orthogonal weight modification, OWM）[8] 算法通过计算与所有的旧数据

张成的特征平面相正交的矩阵 $\boldsymbol{P}$, 在新数据集上训练的过程中, 将梯度的方向按照矩阵 $\boldsymbol{P}$ 的方向修正. 在学习新数据的过程中, 模型权重的更新方向, 与旧输入数据平面正交, 确保新的权重更新不影响从旧数据上学习到的参数. 知识蒸馏连续学习算法 LwF (learning without forgetting)[23] 通过将旧模型作为教师模型在微调过程中约束新模型的输出表示, 使其与旧模型的输出表示接近, 从而降低学习新知识过程中对旧模型输出的改变. 弹性权重固定(elastic weight consolidation, EWC)[9] 算法计算旧任务上模型参数的 Fisher 矩阵, 在微调过程中加载该 Fisher 矩阵来迫使模型保留旧任务上学习到的知识.

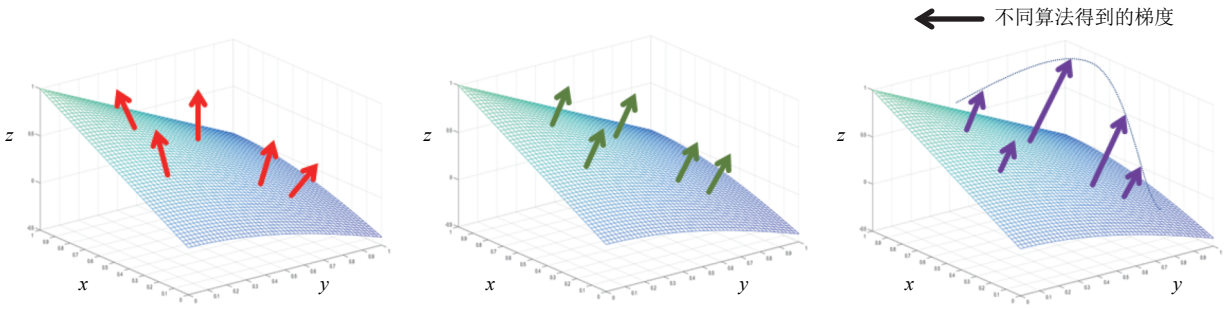
2) 面向跨数据集合成语音检测的方法

如何克服灾难性遗忘一直是连续学习领域研究的一个重点问题. 针对合成语音检测, 通过保存旧数据集的数据, 在新数据集上微调的过程中加入保存的数据联合训练, 以达到降低对旧数据遗忘的目的. 基于连续学习领域经典的经验重放算法, Monteiro 等人[24] 针对合成语音检测提出了一种集成式模型联合训练策略, 相比于仅仅将新旧数据混合训练, 提高了模型的识别准确率; Wang 等人[25] 提出了一种域适应性框架用于新旧数据的混合训练, 提高了模型对新旧数据的递归最小二乘法(recursive least squares, RLS)[26-28]. 然而, 基于经验重放的算法需要保存旧数据集的全部或部分数据, 在多数据集连续学习中, 若每个旧数据集都保存全部或部分数据, 将会对数据存储造成极大的挑战. 另外, 在实际应用中, 多数情况下旧数据集的数据是不可获取的. 例如商业公司训练的模型可能会公开模型, 但极少数会同时公开其训练数据集. 因此, 克服灾难性遗忘更通用的算法是对模型训练中的梯度方向, 模型参数或模型的输出表示进行修正. 例如, Ma 等人[29] 于 2021 年提出的针对合成语音检测的克服灾难性遗忘的知识蒸馏连续表征算法 DFWF(detecting fake without forgetting)[29]. 该算法基于连续学习中经典的正则化算法, 通过分别对全部数据和真实语音计算正则化惩罚项, 在新数据集上微调的过程中约束模型的输出表示, 从而克服灾难性遗忘. 然而, 该算法也存在一定的不足. 由于施加了过强的正则化约束, 限制了模型在新数据集上参数的更新, 导致该算法在新数据集上的识别准确率低于微调算法.

在上述连续学习算法中, OWM 作为一种主流算法被用于机器视觉等领域在多个任务连续学习下克服对旧任务的遗忘, 并取得了显著的成果. 然而, 正交权重修正仅仅着眼于修改模型的权重更新方向, 没有考虑不同权重更新程度的差异. 为了克服新模型对已学知识的灾难性遗忘, 同时提高对新数据的识别能力, 受 OWM 算法的启发, 本文提出了一种不需要获取旧数据, 通过引入额外正则化约束的连续学习算法——弹性正交权重修正(elastic orthogonal weight modification, EOWM). 为了验证所提算法的性能, 本文以跨数据集的合成语音检测为例, 将 2 个常用的合成语音数据集 ASVspoof2019[3] 和 ASVspoof-2015[1] 分别作为旧数据集和新数据集进行了实验. 实验结果采用合成检测常用的等错误率(equal error rate, EER)[1] 作为衡量标准. 实验结果表明, 本文的算法 EOWM 在旧数据集上的 EER 与直接微调相比, EER 从 7.334% 降低至 0.821%, 相对降低了 90% 左右, 在新数据集上的 EER 从 0.513% 降低至 0.315%, 相对降低了 40% 左右; 另外与 EWC[9]、LwF[23]、DFWF[29] 以及 OWM 算法相比, EOWM 算法在新旧数据集上都取得了最好的效果.

1 算法介绍

在合成语音检测领域, 不同数据集之间的模式失配导致在新数据集上微调后的模型存在对旧数据集的灾难性遗忘. 目前 OWM 算法作为一种主流的连续学习中克服灾难性遗忘的算法, 在机器视觉等领域取得了良好的效果. 然而该算法也存在不足, 即 OWM 算法仅仅考虑了连续学习过程中梯度方向的正交化修正, 而没有考虑修正的幅度. 受 OWM 算法的启发, 本文提出了 EOWM 算法. 现有的深度模型的参数往往是海量的, 其中一部分对于旧数据的检测起到了至关重要的作用, 而剩下的一部分权重, 其变化对旧数据集的影响较小[9]. 对于那些对旧数据检测起关键作用的权重参数, 在新任务学习过程中应尽量减少对它们的修改; 而对那些对于旧任务相对不重要的参数, 可以通过重点更新这部分参数来提高模型在新数据集上的检测准确率. 由此, 本文提出了 EOWM 算法, 该算法与经典的 BP 算法和 OWM 算法梯度的对比如图 1 所示. EOWM 算法不仅考虑了模型在新任务上的权重更新方向, 而且通过引入约束项, 针对不同的权重赋予不同的权重更新幅度. 具体来说, 对于那些对旧任务的识别相对重要的参数, EOWM 在正交化其更新方向的同时, 约束其更新程度, 使得这部分参数的变化较小, 从而进一步降低模型对旧任务的遗忘; 对于那些改变后对旧任务的识



(a) BP算法 (b) OWM算法 (c) EOWM算法

Fig. 1 Gradients calculated by different algorithms

图 1 不同算法计算得到的梯度

别影响较小的参数，EOWM 正交化它们的修正方向，同时放松对其更新程度的约束，使得这部分参数可以更好地学习新任务的新知识，从而提高模型对新任务的学习效果.

1.1 权重方向修正

本文的 EOWM 算法中权重方向修正部分受 OWM[8] 算法的启发. 假设给定一个深度模型 $\boldsymbol{M}$，包括 L 个线性前馈层，所有的线性层之间均为相同的激活函数 $g(\cdot)$. 在连续学习任务中，给定一个任务序列包括 J 个任务. 在前一个任务上训练结束的模型，会紧跟着在下一个任务上训练. 在模型的训练过程中，模型可见的数据仅仅是当前任务或当前数据集的数据，而任务序列其他任务的数据均不可见. 在前一个任务上训练完毕的模型称为下一个任务的旧模型. 在该任务序列中，按照如式(1)修正模型的梯度方向：

$$\begin{aligned}&\boldsymbol{W}_l(i,j)=\boldsymbol{W}_l(i-1,j)+\gamma(i,j)\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j),\ \text{when } j=1,\\&\boldsymbol{W}_l(i,j)=\boldsymbol{W}_l(i-1,j)+\gamma(i,j)\boldsymbol{P}_l\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j),\ \text{when } j>1,\end{aligned}\tag{1}$$

其中，i为第i批数据样本；j为任务序列中的第j个任务；$\boldsymbol{W}_l(i,j)$ 为第 l 个线性层在第 j 个任务上的第i批数据样本上的参数矩阵；$\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j)$ 为经典 BP 算法计算的该参数的梯度；$\gamma(i,j)$ 表示模型在第 i 批样本和第 j 个任务下的学习率.

可以看出，在任务序列的第 1 个任务上，EOWM 按照经典 BP 算法更新模型参数，不对梯度方向进行修正；从第 2 个任务开始，为了克服模型对旧知识的遗忘，EOWM 引入了一个梯度方向的修正矩阵 $\boldsymbol{P}_l$，将各层的梯度 $\nabla\boldsymbol{W}_l^{\mathrm{BP}}$ 分别与该修正矩阵 $\boldsymbol{P}_l$ 相乘，从而改变梯度的修正方向，使其与旧任务的数据平面正交. 在此基础上正交化的新任务的权重更新不影响从旧数据中获取到的知识，从而减少对旧知识的遗忘. 由于在连续学习的任务序列中，模型对旧数据是不可见的，所以不能直接利用旧数据计算与数据平面正交的修正矩阵. 因此，EOWM 使用迭代的递推最小二乘(recursive least square, RLS)[26-28]，直接利用当前任务的数据迭代计算修正矩阵 $\boldsymbol{P}_l$，如式(2)所示：

$$\begin{aligned}&\boldsymbol{P}_l(i,j)=\boldsymbol{P}_l(i-1,j)-\boldsymbol{k}_l(i,j)\bar{\boldsymbol{x}}_{l-1}(i,j)^{\mathrm{T}}\boldsymbol{P}_l(i-1,j),\\&\boldsymbol{k}_l(i,j)=\frac{\boldsymbol{P}_l(i-1,j)\bar{\boldsymbol{x}}_{l-1}(i,j)}{\alpha_{i,j}+\bar{\boldsymbol{x}}_{l-1}(i,j)^{\mathrm{T}}\boldsymbol{P}_l(i-1,j)\bar{\boldsymbol{x}}_{l-1}(i,j)},\end{aligned}\tag{2}$$

其中：$\bar{\boldsymbol{x}}_{l-1}(i,j)$和$\bar{\boldsymbol{x}}_{l-1}(i,j)^{\mathrm{T}}$分别表示以输入为第$j$个任务的第$i$批数据样本的均值时，$l-1$层的输出和其转置；$\alpha_{i,j}$是一个超参数，$\alpha_0$为其人工设置的初始值，在我们的实验中初始值$\alpha_0=0.1$，$\alpha_{i,j}$按照式(3)随连续学习的过程逐渐衰减.

$$\alpha_{i,j}=\alpha_0\lambda^{\frac{i}{n_j}},\tag{3}$$

其中：λ 为人工设定的小于 1 的常数；n_j则是第 j个任务下的样本批次总数，即$i=1,2,\cdots,n_j$.

1.2 权重幅度约束

除了修正梯度的方向，EOWM 还通过计算参数对于旧任务的重要性矩阵，对模型的参数按照其对于旧任务的重要性进行约束. 参数对于旧任务的重要性从模型的角度可以认为是模型的不同参数的改变对旧模型识别效果影响的程度. 本文规定，对于那些改变后会极大地降低或提升模型识别准确率的参数被认为是对模型更重要的参数. 从函数的角度，若将模型视为函数，则可以通过函数在不同参数处的导数来衡量该参数对模型整体识别效果的影响. 对于深度模型 $\boldsymbol{M}$，在训练过程中，模型参数的变化估计量如式(4)所示.

$$\boldsymbol{M}(\boldsymbol{x}_l(i,j),\omega+\delta)-\boldsymbol{M}(\boldsymbol{x}_l(i,j),\omega)\approx\boldsymbol{g}_{i,j}\boldsymbol{x}_l(i,j)\delta,\tag{4}$$

其中：ω和δ分别表示模型在某一时刻的特定参数和其一个微小的变化量；$\boldsymbol{g}_{i,j}\boldsymbol{x}_l(i,j)$表示在输入$\boldsymbol{x}_l(i,j)$时，参数$\omega$变化$\delta$的过程中对整个模型参数集的影响，即参数$\omega$对于整体模型 $\boldsymbol{M}$ 的重要性，其计算如式(5)所示：

$$\boldsymbol{g}_{i,j}\boldsymbol{x}_l(i,j)=\frac{\partial \boldsymbol{M}(\boldsymbol{x}_l(i,j))}{\partial \omega}. \tag{5}$$

由上述分析可知，$\boldsymbol{g}_{i,j}(\boldsymbol{x}_l(i,j))$表示了输入一部分数据时该参数对模型的重要性. 在每一个任务训练完毕后，我们把所有的数据都输入旧模型，根据式(6)计算出每个参数对于第 j 个任务中模型从所有输入数据学习到的知识的重要性矩阵$\boldsymbol{\Omega}_j$.

$$\boldsymbol{\Omega}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}\|\boldsymbol{g}_{i,j}(\boldsymbol{x}_l(i,j))\|. \tag{6}$$

对于每个参数的重要性矩阵$\boldsymbol{\Omega}_j$，在新任务上训练过程中，EOWM 额外引入一个二次约束项来根据参数重要性调节模型参数更新的程度，与正交化梯度修正方向结合，在连续学习的过程中，克服灾难性遗忘.

1.3 弹性正交权重修正中的参数更新

本文的算法不仅约束权重的修正方向，还约束了权重的修正幅度. 该算法在连续学习中模型参数的更新如式(7)所示：

$$\boldsymbol{W}_l(i,j)=\boldsymbol{W}_l(i-1,j)+\gamma(i,j)\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j),\ \text{when } j=1,$$
$$\boldsymbol{W}_l(i,j)=\boldsymbol{W}_l(i-1,j)+\gamma(i,j)\boldsymbol{P}_l\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j)+\boldsymbol{R},\ \text{when } j>1.$$

$$\boldsymbol{R}=\frac{\lambda}{2}\sum\boldsymbol{\Omega}_{j-1}(\omega-\omega^*)^2, \tag{7}$$

其中：ω^*为旧模型在第 $j-1$个任务结束后的参数；λ 是一个人工设定的超参数，用来平衡约束程度大小.

从式(7)可以看出，在本文提出的 EOWM 中，首先计算了与旧任务的输入数据平面相正交的修正矩阵$\boldsymbol{P}_l$，通过将该修正矩阵与后向梯度相乘的方式，使得新模型的参数更新与旧任务的数据平面正交，减少了新任务的学习对已学知识的干扰；EOWM 又引入了额外的正则化项$\frac{\lambda}{2}\sum\boldsymbol{\Omega}_{j-1}(\omega-\omega^*)^2$，按照对已学知识的重要性弹性约束模型参数的更新大小，对于那些对旧模型的识别起关键作用的参数，EOWM 减少其在新模型上的变化量，从而进一步保留已学知识的完整. EOWM 算法如算法 1 所示.

算法 1. EOWM.

输入：给定一个连续学习任务序列，包含J个任务，每个任务包含 1 个数据集，所有的数据集的数据集合为X，数据集合中的数据样本为 $\boldsymbol{x}$，随机初始化包含L层的深度模型 $\boldsymbol{M}$；初始化超参数为α_0和γ，初始化学习率为 γ，初始化第 1 层的梯度方向修正矩阵$\boldsymbol{P}_l$为单位对角矩阵；

输出：训练完毕的模型 $\boldsymbol{M}$.

① for j in J do
② if j=1 then
③ for $\boldsymbol{x}(i,j)$ in X_j do
④ $\boldsymbol{W}_l(i,j)=\boldsymbol{W}_l(i-1,j)\gamma(i,j)\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j)$;
⑤ end for
⑥ for $\boldsymbol{x}(i,j)$ in X_j do
⑦ $\boldsymbol{g}_{i,j}\boldsymbol{x}_l(i,j)=\frac{\partial \boldsymbol{M}(\boldsymbol{x}_l(i,j))}{\partial \omega}$;
⑧ $\boldsymbol{\Omega}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}\|\boldsymbol{g}_{i,j}(\boldsymbol{x}_l(i,j))\|$;
⑨ end for
⑩ else
⑪ for $\boldsymbol{x}(i,j)$ in X_j do
⑫ $\alpha_{i,j}=\alpha_0\lambda^{\frac{i}{n_j}}$;
⑬ $\boldsymbol{P}_l(i,j)=\boldsymbol{P}_l(i-1,j)-\boldsymbol{k}_l(i,j)\bar{\boldsymbol{x}}_{l-1}(i,j)^{\mathrm{T}}\boldsymbol{P}_l(i-1,j)$;
⑭ $\boldsymbol{k}_l(i,j)=\frac{\boldsymbol{P}_l(i-1,j)\bar{\boldsymbol{x}}_{l-1}(i,j)}{\alpha_{i,j}+\bar{\boldsymbol{x}}_{l-1}(i,j)^{\mathrm{T}}\boldsymbol{P}_l(i-1,j)\bar{\boldsymbol{x}}_{l-1}(i,j)}$;
⑮ $\boldsymbol{W}_l(i,j)=\boldsymbol{W}_l(i-1,j)+\gamma(i,j)\boldsymbol{P}_l\nabla\boldsymbol{W}_l^{\mathrm{BP}}(i,j)+\boldsymbol{R}$;
⑯ $\boldsymbol{R}=\frac{\lambda}{2}\sum\boldsymbol{\Omega}_{j-1}(\omega-\omega^*)^2$;
⑰ end for
⑱ for $\boldsymbol{x}(i,j)$ in X_j do
⑲ $\boldsymbol{g}_{i,j}\boldsymbol{x}_l(i,j)=\frac{\partial \boldsymbol{M}(\boldsymbol{x}_l(i,j))}{\partial \omega}$;
⑳ $\boldsymbol{\Omega}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}\|\boldsymbol{g}_{i,j}(\boldsymbol{x}_l(i,j))\|$;
㉑ end for
㉒ end if
㉓ end for

总的来说，EOWM 通过有选择性地正交化更新模型参数，来使模型拟合到新任务上的同时，减少拟合过程对旧任务的干扰. 由于深度模型的参数空间是低秩的[26]，部分参数在优化过程中对模型的性能都没有显著的影响. 因此 EOWM 通过训练模型的一部分参数，使得模型在一个新的任务上达到良好的性能. 具体来说，EOWM 通过引入正则化项，使得模型在新任务上学习时减少对旧任务重要参数的修正，改变那些对旧任务不重要的参数. 由于模型的参数空间是严重不满秩的，所以只重点更新对旧任务不重要的参数，可以做到让模型拟合到新任务上，从而显著提高模型在新数据集上的检测率.

2 实验及结果分析

为了验证本文的算法在连续学习中，分别对于

克服灾难性遗忘和获取新知识的效果，本文针对合成语音检测任务设置了一系列实验. 实验在连续学习中常见的也是最基本的旧数据集到新数据集的迁移过程，比较不同模型在旧数据集和新数据集上的EER[1]. 本文首先使用旧数据集训练一个基线系统，然后将该基线系统在旧数据集和新数据集上分别测试其检测效果，最后将基线系统迁移到新数据集上. 在本轮所有的实验中，每个模型或算法均运行 7 次以上，最后的结果是所有结果的均值.

2.1 基本定义

本文实验的数据集采用合成语音检测领域常用的 2 个国际竞赛数据集，分别是 ASVspoof2015[1] 和及 ASVspoof2019LA[3]. 其中 ASVspoof2019LA 是 ASVspoof2019 的子集，包括 30 名男性说话人和 37 名女性说话人的语音. 将其划分成 3 个子数据集，分别是训练集、验证集和测试集. 其中训练集和验证集的合成攻击方式相同，包括 4 种语音合成算法 TTS(text-to-speech)和 2 种语音转换(voice conversion, VC)算法. 其中，真实语音是从 VCTK 语料库[30] 中选择的. 与训练集和验证集不同，测试集包含不同的合成语音攻击算法，目的是提高模型对未知类型的合成语音的鉴别能力，从而提高模型的泛化性. ASVspoof2015 中包含 45 名男性说话人和 61 名女性说话人的语音，所有的语音中均在无噪声的环境下录制，避免了噪声对模型识别的影响. 本文约定，旧数据集为 ASVspoof-2019LA，新数据集为 ASVspoof2015；旧模型为旧数据集上训练完毕的模型，新模型为旧模型在新数据集上微调或连续学习后的模型. 相关数据集细节如表 1 所示. 其中 A01~A19 为 ASVspoof2019 数据集中使用的语音合成算法，包括语音合成和语音转换以及二者的结合[3]；S01~S10 为 ASVspoof2015 数据集中使用的语音合成算法[1].

Table 1 Audio Number and Type of Fake Language in Experimental Datasets

表 1 实验数据集中的语音数目和合成语音类型

数据集	ASVspoof2019LA		
	真实语音	合成语音	合成类型
训练集	2 580	22 800	A01~A06
验证集	2 548	22 296	A01~A06
测试集	7 355	63 882	A07~A19
数据集	ASVspoof2015		
	真实语音	合成语音	合成类型
训练集	3 750	12 625	S01~S05
验证集	3 497	49 875	S01~S05
测试集	9 404	184 000	S01~S10

2.2 实验配置

1)特征. 目前基于预训练模型的合成语音检测模型已经证明其具有最低的 EER[31]，因此本文使用预训练模型 Wav2vec 2.0[32] 作为本实验的特征提取器. Wav2vec 2.0 的参数从预训练模型 XLSR-53[33] 中加载. 所有的特征均来自预训练模型编码器的末端，隐藏层维度为 1 024.

2)模型. 本文使用合成语音检测领域常用的自注意力卷积神经网络(self attention convolutional neural network, S-CNN)作为分类器. 按照前向推理过程，分类器 S-CNN 包含 3 个 1D 卷积层、1 个自注意力层和 2 个全连接层. 第 1 卷积层的输入维度为 256，所有卷积层的隐藏层维度为 80. 内核大小和步长分别设置为 5 和 1. 所有连接层的隐藏层维度为 80，最后一层的输出维度为 2.

3)训练细节. 本实验包含 2 个过程，分别是在 ASVspoof2019LAs 上训练得到的旧模型和旧模型从 ASVspoof2019 到 ASVspoof2015 上迁移. 在这 2 个过程中，本文微调了模型权重，包括预训练模型 XLSR-53 和分类器 S-CNN. 所有参数都由 Adam 优化器训练，批大小为 2，学习率 $\gamma = 0.000\,1$. 对于所有的卷积层，弹性正交修正中的 α 初始化为 0.000 01，对于自注意力层，α 初始化为 0.000 1，对于全连接层，α 初始化为 0.1.

本文首先在 ASVspoo2019LA 数据集的训练集上训练识别模型得到了旧模型，然后分别在 ASVspoof-2019LA 和 ASVspoof2015 数据集上验证了旧模型的检测能力. 表 2 显示了在旧数据集和新数据集的测试集上的基线检测性能. 本文的基线在 ASVspoof2019 上的 EER 与目前最好的结果非常接近[9]. 尽管该模型在 ASVspoo2019LA 上取得了令人满意的性能，但其

Table 2 EER on the Old Dataset ASVspoof2019LA and New Dataset ASVspoof2015

表 2 旧数据集 ASVspoof2019LA 和新数据集 ASVspoof2015 上的等错误率

模型	训练步数	测试集	
		ASVspoof2019LA	ASVspoof2015
模型-1	300	0.037 51	0.063 16
模型-2	400	0.029 75	0.085 17
模型-3	500	0.017 94	0.099 88
模型-4（基线系统）	800	0.002 58	0.245 32
模型-5	900	0.002 59	0.256 98
模型-6	1 000	0.002 62	0.278 72

检测精度在新数据集上显著下降，这验证了前文的论点，即模型在跨数据集场景下识别精度会有较大幅度的降低. 除此之外，表 2 中还给出了不同训练步数的结果，其中模型-1 到模型-6 是本文使用的模型通过逐渐增加训练步数得到的旧模型. 从表 2 中可以看出随着在旧数据集上训练步数的增加，模型在旧数据集上的识别精度逐渐提高，但在新数据集上的识别精度却逐渐降低.

2.3 弹性系数λ对模型的影响

本文还针对提出的 EOWM 算法中不同的弹性系数进行了实验，来验证该超参数对模型克服遗忘和获取新知识的影响，结果如表 3 所示. 可以观察到，在弹性系数λ位于 0.5 左右时模型在新旧 2 个数据集上都达到了最低的 EER. 在$\lambda<0.5$的范围内，随着λ的不断减小，在旧数据集上的 EER 有明显降低，在新数据集上的变化则相对较小. 而在$\lambda<0.5$ 的范围内，该变化趋势相对平滑. 这表明，在非最优解的情况下，较大的λ相对较小的λ，在连续学习的新旧数据集上更有优势.

Table 3 EER of Some Models with Various Elastic Coefficients λ on Old Dataset ASVspoof2019LA and New Dataset ASVspoof2015

表 3 带有不同弹性系数λ的弹性正交权重修正在旧数据集 ASVspoof2019LA 和新数据集 ASVspoof2015 上的等错误率

弹性系数 λ	测试集	
	ASVspoof2019LA	ASVspoof2015
λ=0.1	0.016 77	0.005 11
λ=0.1	0.014 32	0.005 11
λ=0.3	0.009 76	0.004 65
λ=0.4	0.008 22	0.003 16
λ=0.5	0.008 21	0.003 15
λ=0.6	0.008 36	0.003 18
λ=0.7	0.008 67	0.003 79
λ=0.8	0.009 56	0.004 52
λ=0.9	0.009 77	0.004 73
λ=1.0	0.009 89	0.005 16
基线系统	0.025 80	0.245 32

2.4 EOWM 与其他算法的比较

本文将所提的算法 EOWM 与其他几个连续学习领域的主流算法进行了对比，结果如表 4 所示，其中 EWC, OWM, LwF, DFWF 的结果均为本文复现的原论文中的算法得到的结果；EOWM 为本文提出的算法得到的结果. 其中联合训练表示在新数据集上训练的过程中，加入全部的旧数据集的训练数据一起训练，这个结果表示了连续学习算法在新旧数据集上的 EER 下限. 从结果可以看出，所有的连续学习算法与直接微调相比，在旧数据集上的 EER 均有所降低，其中 EOWM 克服遗忘的效果最为明显，是表中所有算法中唯一一个在旧数据集上达到了低于 1 的 EER 的算法. 与直接微调相比，仅有 LwF 和 EOWM 在新数据集上达到了更高的识别准确率，其中本文的算法达到了最低的 EER. 相比于 EER 下限的联合训练算法，EOWM 在新数据集上达到了相似的结果，这表明 EOWM 在连续学习序列中对新数据的训练有很好的效果，并且不需要引入旧的数据.

Table 4 EER of Different Continual Learning Algorithms on the Old Dataset ASVspoof2019LA and New Dataset ASVspoof2015

表 4 不同的连续学习算法在旧数据集 ASVspoof2019LA 和新数据集 ASVspoof2015 上的等错误率

模型	是否需要旧数据	是否需要新数据	测试集	
			ASVspoof2019LA	ASVspoof2015
基线系统	是	否	0.002 58	0.245 32
联合训练	是	是	0.004 06	0.002 01
Fine-tune	否	是	0.073 34	0.005 13
EWC	否	是	0.028 42	0.005 15
OWM	否	是	0.024 88	0.005 15
LwF	否	是	0.032 17	0.003 33
DFWF	否	是	0.018 53	0.006 79
EOWM（本文）	否	是	0.008 21	0.003 15

2.5 EOWM 在新旧任务差异较大时的实验

为了验证 EOWM 在区分度更大的新旧任务上的性能，我们还设置了在 ASVspoof2019LA、In-the-Wild[34]、ASVspoof2019LA 和 VCC2020[35] 数据集上，EOWM 与其他算法的对比. 其中 In-the-Wild 数据集为在实际环境中搜集的音频，具有一定的环境噪声；VCC2020 是一个多语种数据集，包括英语、中文、德语等. 表 5 和表 6 的实验结果表明，EOWM 在新旧任务的伪造音频特征差异较大时，相比于其他算法，仍具有最好的学习和克服灾难性遗忘的效果.

2.6 EOWM 的消融实验

本文对所提的算法进行了消融实验，结果如表 7 所示. 可以观察到在去掉权重修正约束后，模型在新旧数据集上的 EER 均有较大的上升；去掉权重方向修正后，模型在新数据集上的识别效果基本不变，在

Table 5 EER of Different Continual Learning Algorithms on the Old Dataset ASVspoof2019LA and New Dataset In-the-Wild

表 5 不同的连续学习算法在旧数据集 ASVspoof2019LA 和新数据集 In-the-Wild 上的等错误率

模型	是否需要旧数据	是否需要新数据	测试集	
			ASVspoof2019LA	In-the-Wild
基线系统	是	否	0.002 58	0.245 32
联合训练	是	是	0.027 40	0.021 60
Fine-tune	否	是	0.209 76	0.049 78
EWC	否	是	0.080 39	0.056 15
OWM	否	是	0.081 30	0.050 65
LwF	否	是	0.064 53	0.049 98
DFWF	否	是	0.063 24	0.062 75
EOWM（本文）	否	是	0.057 49	0.038 95

Table 6 EER of Different Continual Learning Algorithms on the Old Dataset ASVspoof2019LA and New Dataset VCC2020

表 6 不同的连续学习算法在旧数据集 ASVspoof2019LA 和新数据集 VCC2020 上的等错误率

模型	是否需要旧数据	是否需要新数据	测试集	
			ASVspoof2019LA	VCC2020
Baseline	是	否	0.002 58	0.245 32
联合训练	是	是	0.009 65	0.024 98
Fine-tune	否	是	0.087 55	0.056 47
EWC	否	是	0.034 94	0.062 89
OWM	否	是	0.030 86	0.064 32
LwF	否	是	0.042 34	0.045 66
DFWF	否	是	0.018 74	0.073 55
EOWM（本文）	否	是	0.015 36	0.039 94

Table 7 Ablation Study of EOWM on Old Dataset ASVspoof2019LA and New Dataset ASVspoof2015

表 7 EOWM 在旧数据集 ASVspoof2019LA 和新数据集 ASVspoof2015 上的消融实验

模型	测试集	
	ASVspoof2019LA	ASVspoof2015
基线系统	0.258	24.532
联合训练	**0.406**	**0.201**
EOWM	**0.821**	**0.315**
权重修正约束	2.488	0.515
权重方向修正	7.344	0.513

注：黑体数值表示最优值.

旧数据集上的 EER 大幅度提高. 这表明权重修正约束对获取新知识和克服遗忘均有积极影响；而权重方向修正对获取新知识几乎没有影响，而对克服遗忘有积极影响. 相对来说，权重修正约束侧重于帮助模型获取新知识的同时不影响已学知识，而权重方向修正更侧重于尽可能保护已学知识.

3 结 论

本文提出了一种连续学习算法——弹性正交权重修正算法 EOWM 用于克服灾难性遗忘. 该算法与基于梯度修正的连续学习算法类似，不需要获取旧任务的数据，且适用于大部分的学习场景. 与正交权重修正算法相比，EOWM 不仅考虑了在新任务上对模型权重更新方向的修正，还引入了对模型权重修正幅度的约束. 具体来说，对于那些对旧任务起重要作用的参数，我们减少它们在新任务上的变化幅度. 实验表明，本文的算法在合成语音检测方面优于多种主流的连续学习算法，该算法引入的对模型权重修正程度的约束对新旧任务的识别都产生了明显的积极影响. 除此以外，本文通过实验给出了所提算法中部分参数设置的建议. 未来的改进可以着眼于进一步探索针对模型参数的修正策略，根据不同类型的参数，针对性地修正其方向和大小，来进一步克服遗忘和提高模型对新知识的学习能力.

作者贡献声明：张晓辉提出了算法思路、完成实验并撰写论文；易江燕提出指导意见并修改论文；陶建华提供了实验条件；周俊佐修改了论文.

参 考 文 献

[1] Wu Zhizheng, Kinnunen T, Evans N, et al. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge[C]//Proc of the 16th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2015: 2037−2041

[2] Kinnunen T, Sahidullah M, Delgado H, et al. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection[C]//Proc of the 18th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 2−6

[3] Todisco M, Wang Xin, Vestman V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection[C]//Proc of the 20th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2019: 1008−1012

[4] Yamagishi J, Wang Xin, Todisco M, et al. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection[J]. arXiv preprint, arXiv: 2109.00537, 2021

[5] Yi Jiangyan, Fu Ruibo, Tao Jianhua, et al. ADD 2022: The first audio deep synthesis detection challenge[C]//Proc of 47th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2022: 9216−9220

[6] Müller N, Czempin P, Dieckmann F, et al. Does audio deepfake detection generalize[C]//Proc of the 23rd Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2022: 2783−2787

[7] Zhang You, Zhu Ge, Jiang Fei, et al. An empirical study on channel effects for synthetic voice spoofing countermeasure systems[C]//Proc of the 22nd Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2021: 4309−4313

[8] Zeng Guanxiong, Chen Yang, Cui Bo, et al. Continual learning of context-dependent processing in neural networks[J]. Nature Machine Intelligence, 2019, 1: 364−372

[9] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521−3526

[10] Parisi G, Kemker R, Part J, et al. Continual lifelong learning with neural networks: A review[J]. Neural Networks, 2019, 113: 54−71

[11] Lopez-Paz D, Muandet K, Schölkopf B, et al. Towards a learning theory of cause-effect inference[J]. arXiv preprint, arXiv: 1502.02398, 2015

[12] Aljundi R, Rohrbach M, Tuytelaars T. Selfless Sequential Learning[J]. arXiv preprint, arXiv: 1806.05421, 2019

[13] Lomonaco V, Maltoni D. CORe50: A new dataset and benchmark for continuous object recognition[J]. arXiv preprint, arXiv: 1705.03550, 2017

[14] Shmelkov K, Schmid C, Alahari K, et al. Incremental learning of object detectors without catastrophic forgetting[C]//Proc of the 16th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 3420−3429

[15] Rebuffi S, Kolesnikov A, Sperl G, et al. ICaRL: Incremental classifier and representation learning[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 5533−5542

[16] Rusu A, Rabinowitz N, Desjardins G, et al. Progressive neural networks[J]. arXiv preprint, arXiv: 1606.04671, 2016

[17] Schwarz J, Czarnecki W, Luketina J, et al. Progress & compress: A scalable framework for continual learning[J]. arXiv preprint, arXiv: 1805.06370, 2018

[18] Yoon J, Yang E, Lee J, et al. Lifelong learning with dynamically expandable networks[J]. arXiv preprint, arXiv: 1708.01547, 2018

[19] Lopez-Paz D, Ranzato M. Gradient episodic memory for continual learning[C]//Proc of the 31st Annual Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2017: 6467−6476

[20] Castro F, Marín-Jiménez M, Guil N, et al. End-to-end incremental learning[C]//Proc of the 15th European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 233−248

[21] Wu Yue, Chen Yinpeng, Wang Lijuan, et al. Large-scale incremental learning[C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 374−382

[22] Lee K, Lee K, Shin J, et al. Overcoming catastrophic forgetting with unlabeled data in the wild[C]//Proc of the 17th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 312−321

[23] Li Zhizhong, Hoiem D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2935−2947

[24] Monteiro J, Alam J, Falk T, et al. An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers[C]//Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 6599−6603

[25] Wang Hongji, Dinkel H, Wang Shuai, et al. Dual-adversarial domain adaptation for generalized replay attack detection[C]//Proc of the 21st Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2020: 1086−1090

[26] Haykin S. Adaptive Filter Theory[M]. London: Pearson, 2014

[27] Singhal S, Wu L. Training feed-forward networks with the extended Kalman algorithm[C]//Proc of the 47th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 1989: 1187−1190

[28] Shah S, Palmieri F, Datum M. Optimal filtering algorithms for fast learning in feedforward neural networks[J]. Neural Network, 1992, 5(5): 779−787

[29] Ma Haoxin, Yi Jiangyan, Tao Jianhua, et al. Continual learning for fake audio detection[C]//Proc of the 22nd Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2021: 886−890

[30] Veaux C, Yamagishi J, MacDonald K, et al. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit[EB/OL]. (2019-11-13)[2023-04-10]. https://datashare.ed.ac.uk/handle/10283/3443

[31] Wang Xin, Yamagishi J. Investigating self-supervised front ends for speech spoofing countermeasures[J]. arXiv preprint, arXiv: 2111.07725, 2021

[32] Baevski A, Zhou Yuhao, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[C]//Proc of the 34th Annual Conf on Neural Information Processing Systems (NIPS). Cambridge, MA: MIT, 2020: 12449−12460

[33] Conneau A, Baevski A, Collobert R, et al. Unsupervised cross-lingual representation learning for speech recognition[C]//Proc of the 22nd Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2021: 2426−2430

[34] Muller N, Czempin P, Dieckmann F, et al. Does audio deepfake detection generalize[C]//Proc of the 23rd Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2022: 2783−2787

[35] Zhao Yi, Wen-Chin H, Tian Xiaohai, et al. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion[J]. arXiv preprint, arXiv: 2008.12527, 2020



Zhang Xiaohui, born in 1998. Master candidate. His main research interests include deep fake audio detection and continual learning.

张晓辉, 1998 年生. 硕士研究生. 主要研究方向为深度合成音频鉴别、连续学习.



Yi Jiangyan, born in 1984. PhD, Master supervisor. Her main research interests include speech information processing, speech generation and identification, and continuous learning.

易江燕, 1984 年生. 博士, 硕士生导师. 主要研究方向为语音信息处理、语音生成与鉴别、持续学习.



Tao Jianhua, born in 1972. PhD, PhD supervisor. His main research interests include intelligent information fusion and processing, speech processing, affective computing, and big data analysis.

陶建华, 1972 年生. 博士, 博士生导师. 主要研究方向为智能信息融合与处理、语音处理、情感计算、大数据分析.



Zhou Junzuo, born in 2000. Master candidate. His main research interest includes text-to-speech.

周俊佐, 2000 年生. 硕士研究生. 主要研究方向为文本语音转换.