Do You Remember? Overcoming Catastrophic Forgetting for Fake Audio Detection

Xiaohui Zhang [1 2] **Jiangyan Yi** [1] **Jianhua Tao** [3] **Chenglong Wang** [1 4] **Chuyuan Zhang** [1]

Abstract

Current fake audio detection algorithms have achieved promising performances on most datasets. However, their performance may be significantly degraded when dealing with audio of a different dataset. The orthogonal weight modification to overcome catastrophic forgetting does not consider the similarity of genuine audio across different datasets. To overcome this limitation, we propose a continual learning algorithm for fake audio detection to overcome catastrophic forgetting, called Regularized Adaptive Weight Modification (RAWM). When fine-tuning a detection network, our approach adaptively computes the direction of weight modification according to the ratio of genuine utterances and fake utterances. The adaptive modification direction ensures the network can effectively detect fake audio on the new dataset while preserving its knowledge of old model, thus mitigating catastrophic forgetting. In addition, genuine audio collected from quite different acoustic conditions may skew their feature distribution, so we introduce a regularization constraint to force the network to remember the old distribution in this regard. Our method can easily be generalized to related fields, like speech emotion recognition. We also evaluate our approach across multiple datasets and obtain a significant performance improvement on cross-dataset experiments.

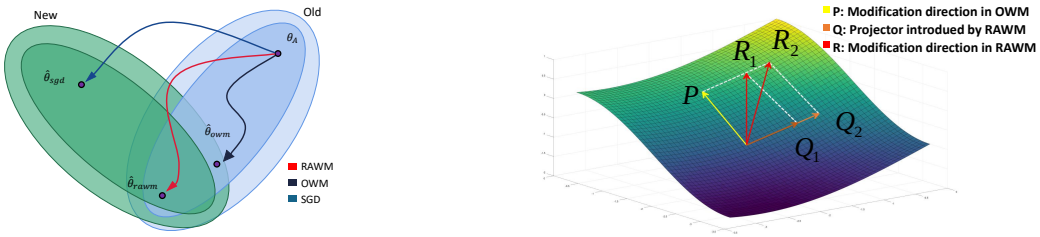
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China [2]School of Computer and Information Technology, University of Beijing Jiaotong, Beijing, China [3]Department of Automation, Tsinghua University, Beijing, China [4]University of Science and Technology of China, Beijing, China. Correspondence to: Jiangyan Yi <jiangyan.yi@nlpr.ia.ac.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

With the development of speech synthesis and voice conversion technology (Wang et al., 2018; 2021), the models can generate human-like speech, which makes it difficult for most people to distinguish the generated audio from the real one. Although this technology has brought great convenience to human life, it has also brought great safety hazards to the country and society. Therefore, fake audio detection has attracted increasing attention in recent years. A series of challenges have been organized to detect fake audio, such as the ASVspoof challenge (Wu et al., 2015; Kinnunen et al., 2017; Todisco et al., 2019; Yamagishi et al., 2021) and the Audio Deep Synthesis Detection (ADD) challenge (Yi et al., 2022). In these competitions, deep neural networks have achieved great success. Currently, large-scale pre-trained models have gradually been applied to fake audio detection and achieved state-of-the-art results on several public fake audio detection datasets (Tak et al., 2022; Martín-Doñas & Álvarez, 2022; Lv et al., 2022; Wang & Yamagishi, 2021). Although fake audio detection achieves promising performance, it may be significantly degraded when dealing with audio of another dataset. The diversity of audio proposes a significant challenge to fake audio detection across datasets (Zhang et al., 2021b;a).

Some approaches have been proposed to improve detection performance across datasets. An ensemble learning method is proposed to improve the detection ability of the model for unseen audio (Monteiro et al., 2020) and a dual-adversarial domain adaptive network (DDAN) is designed to learn more generalized features for different datasets (Wang et al., 2020). Both methods require some audio from the old dataset, but in some practical situations, it is almost impossible to obtain them. For instance, a pre-trained model proposed by a company has been released to the public. It is unfeasible for the public to fine-tune it using the data belonging to the original company. In addition, a data augmentation method is proposed to extract more robust features for detection across datasets (Zhang et al., 2021b), which is only suitable for the datasets with similar feature distribution. In continual learning, a method called Detecting Fake Without Forgetting (DFWF) is proposed for fake audio detection (Ma et al., 2021). Although the above meth-



(a) (b)

Figure 1: Schematic of SGD, OWM, and RAWM. **(a)**, With RAWM, the optimization process searches for configurations that lead to great performance on both old (blue area) and new (green area) datasets. The center parts of the two areas represent better recognition performance than the other, and can be regarded as subspaces of the area mentioned by the OWM. A successful optimized configuration $\hat{\theta}_{rawm}$ stops inside the overlapping subspace. However, the configuration $\hat{\theta}_{sgd}$ obtained by SGD is optimized without considering forgetting, and the configuration $\hat{\theta}_{owm}$ obtained by orthogonal weight modification can reach the overlapping area but not the overlapping subspace. **(b)**, the RAWM adaptively modifies weight direction by introducing a projector that is orthogonal to the projector $\boldsymbol{P}$ proposed by OWM.

ods are effective, they still have some limitations, like the acquisition of old data in the ensemble learning method and the DDAN and deteriorating learning performance in the DFWF. This paper, however, aims to overcome catastrophic forgetting while exerting a positive influence on acquiring new knowledge without any previous samples.

Most fake audio detection datasets are under clean conditions, where the genuine audio has a more similar feature distribution than the fake audio (Ma et al., 2021). A few datasets, however, are collected under different acoustic conditions (Müller et al., 2022), which makes a difference in their feature distributions of genuine audio (Ma et al., 2022). If we modify the model weights as the orthogonal weight modification (OWM) method (Zeng et al., 2019) which introduces a new weight direction orthogonal to all old data, most genuine audio with similar feature distribution across datasets can not be trained effectively. The reason is that new genuine audio is supposed by the OWM to damage learned knowledge, so it modifies new weight direction orthogonal to the old one regardless the new and old genuine audio have similar feature distribution and they can be seen as a whole from the same dataset. Based on the above inference, it is more effective for genuine audio on new datasets to be trained with a direction close to the previous one, rather than orthogonal to it. To address these issues, we propose a continual learning approach, named Regularized Adaptive Weight Modification (RAWM). In our method, if the proportion of fake audio is larger, the modified direction is closer to the orthogonal projector of the subspace spanned by all old input; if the proportion of genuine audio is larger, the modification is closer to the old input subspace. However, old and new datasets are collected from different acoustic conditions in some cases, where genuine audio may have quite different feature distributions. We address this issue by introducing a regularization constraint. This constraint forces the model to remember the old feature distribution. In addition, compared with the experience-replay-based method, RAWM does not require old data, which makes it suitable for most situations. The optimization process of RAWM is compared with that of the Stochastic Gradient Descent search (SGD) and OWM in Figure 1a.

Contributions: We propose a regularized adaptive weight modification algorithm to overcome catastrophic forgetting. There are two essential steps in our method: adaptive weight modification (AWM) and regularization. The former AWM is proposed for continual learning in most situations where genuine audio has similar feature distribution and the latter regularization is introduced to ease the problem that genuine audio may have different feature distribution in a few cases. Although our method is inspired by the feature distribution similarity in fake audio detection, it can also be used in other related tasks, such as speech emotion recognition. The experimental results show that our proposed method outperforms several continual learning methods in acquiring new knowledge and overcoming forgetting, including Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), Learning without Forgetting (LwF) (Li & Hoiem, 2017), OWM, and DFWF. The code of our method has been released in Regularized Adaptive Weight Modification.

2. Related Work

In continual learning, overcoming catastrophic forgetting methods can be divided into the following categories. The regularization methods perform a regularization on the objection function or regulate important weights that are essential for previous tasks (Kinnunen et al., 2017; Zenke et al., 2017b; Aljundi et al., 2018; 2019; Mallya & Lazebnik, 2018; Serra et al., 2018). The dynamic architecture methods reserve their previous knowledge by introducing additional layers or nodes and grow model architecture (Rusu

et al., 2016; Schwarz et al., 2018); (Yoon et al., 2017). The memory-based methods remember their previous data to prevent gradient updates from damage on their learned knowledge. (Lopez-Paz & Ranzato, 2017; Castro et al., 2018; Wu et al., 2019; Lee et al., 2019). The natural gradient descent methods approximate the Fisher information matrix in EWC using the generalized Gauss-Newton method to fast gradient descent (Tseran et al., 2018; Chen et al., 2019).

Although the mainstream continual learning methods, such as the EWC, LwF and OWM, have achieved great success in many fields including image classification (Zeng et al., 2019; Kirkpatrick et al., 2017), object detection (Perez-Rua et al., 2020), semantic segmentation (Cermelli et al., 2020), lifelong language learning (de Masson D'Autume et al., 2019) and sentence representation (Liu et al., 2019). However, the approximation of regularization methods will produce error accumulation in continual learning (Zenke et al., 2017a; Huszár, 2017; Ma et al., 2021). In contrast, our proposed method only needs the current inputs, which leads to a better performance than others in error accumulation. Besides, we relax the regularized constraint in the DFWF and introduce a direction modification to solve the deteriorating learning performance problem.

3. Background

3.1. Orthogonal Weight Modification

The OWM algorithm overcomes catastrophic forgetting by modifying the direction of weights on the new task. The modified direction $\boldsymbol{P}$, which is a square matrix, is orthogonal to the subspace spanned by all inputs of the previous task. The orthogonal projector is constructed by an iterative method similar to the Recursive Least Square (RLS) algorithm (Shah et al., 1992), which hardly requires any previous samples.

We consider a feed-forward network consisting of $L+1$ layers, indexed by $l = 0, 1, \cdots, L$ with the same activation function $g(\cdot)$. The $\overline{\mathbf{x}}_l(i,j) \in \mathbb{R}^s$ represents the output of the lth layer in response to the mean of the ith batch inputs on jth dataset, and the $\overline{\mathbf{x}}_l(i,j)^T$ is the transpose matrix of the $\overline{\mathbf{x}}_l(i,j)$. The modified direction $\boldsymbol{P}$ can be calculated as:

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

where α is a hyperparameter decaying with the number of tasks.

3.2. Learning without Forgetting

The LwF algorithm is inspired by the idea of model distillation, where old knowledge is viewed as a penalty term to regulate the new model representation similar to the old. The model trained on old datasets is replicated into two models with the same parameters. The two models are named teacher and student models in the LwF. In process of training on new datasets, the parameters of the teacher model are frozen to produce its features as "soft labels". The student model is trained by the loss function as:

$$\boldsymbol{L}_{lwf} = \lambda_0 \boldsymbol{L}_{old}(\mathrm{y_o}, \hat{\mathrm{y}}_o) + \boldsymbol{L}_{new}(\mathrm{y_n}, \hat{\mathrm{y}}_n) \tag{2}$$

where λ_0 is a ratio coefficient representing the importance of learned knowledge; $\mathrm{y_o}$ is the "soft label" produced by the teacher model and $\mathrm{y_n}$ is the ground truth of new data; Both $\hat{\mathrm{y}}_o$ and $\hat{\mathrm{y}}_n$ are the softmax output of the student model. Both $\boldsymbol{L}_{old}$ and $\boldsymbol{L}_{new}$ are cross-entropy loss. The former $\boldsymbol{L}_{old}$ regulates the output probabilities $\hat{\mathrm{y}}_o$ to be close to the recorded output $\mathrm{y_o}$ from the teacher model and the latter $\boldsymbol{L}_{new}$ encourages predictions $\hat{\mathrm{y}}_n$ to be consistent with the ground truth $\mathrm{y_n}$.

4. Proposed Method

On most fake audio detection datasets, under the same acoustic conditions, feature distributions of genuine audio are relatively more concentrated than the fake, which means the feature distribution of genuine audio has a smaller variance than that of fake audio (Ma et al., 2021; Yan et al., 2022). Besides, there are also a few datasets whose genuine audio has quite different feature distributions from others (Ma et al., 2022; Müller et al., 2022).

Based on the observations, we propose a continual learning method, named Regularized Adaptive Weight Modification (RAWM), to overcome catastrophic forgetting. There are two essential steps in our method: adaptive direction modification (AWM) and regularization. The AWM is proposed for most situations where genuine audio has similar feature distribution. As shown in Figure 1b, by introducing an extra projector, which is a square matrix orthogonal to the projector proposed by the OWM, our method could adaptively modify weight direction closer to the previous inputs subspace. As for those genuine audio collected from quite different acoustic conditions, it is detrimental for learned knowledge to modify weight according to the rule we mentioned above, because their feature distribution is distinct from others. To address this issue, we introduce a regularization term to force the new distribution of inference to be similar to the old one. Our method does not require any replay of previous samples. In addition, our method is inspired by fake audio detection but it can easily be generalized to other related tasks. The reason is that most of them have one or more classes, like neutral emotion in speech emotion recognition (SER) (Sharma, 2022), with relatively similar feature distribution between different datasets. We also take SER as an example to present how our method generalizes to other fields in Sec. 4.3 and show the process of our algorithm in Algorithm 1.

4.1. Adaptive Weight Modification

We start by introducing an adaptive modification of weight direction according to the ratio β of classes with similar feature distribution between different datasets and others in batch data, which is essential for sequence training on multi-datasets. We first consider a feed-forward network like that described in Sec. 3.1. Then, we introduce a square matrix $\boldsymbol{Q}$ as a projector that is orthogonal to the $\boldsymbol{P}$ proposed by the OWM algorithm. This orthogonal projector can be written as Eq 3:

$$\boldsymbol{Q} = \beta[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}] \tag{3}$$

where the projector $\boldsymbol{P}$, which is orthogonal to the subspace spanned by all previous inputs, can be calculated as Eq 1 and $\boldsymbol{I}$ is an identity matrix. The construction of the orthogonal projector $\boldsymbol{Q}$ is mathematically sound (Haykin, 2002; Ben-Israel & Greville, 2003; Bengio & LeCun, 2007). To verify the modification direction according to the essential ratio β, we introduce the β defined as:

$$\beta = \frac{\sum_{t=1}^{b} N_t + 1}{\sum_{t=b+1}^{b+c} N_t + 1} \tag{4}$$

in which the $N_t, t \in [1, b]$ represents the number of batch samples of b classes with relatively similar feature distributions on old and new datasets, respectively; the $N_t, t \in [b+1, c]$ represents the number of batch samples of other c classes. By adding one to both the numerator and denominator, β can be calculated when all the batch data belong to classes in the numerator. As illustrated in Eq 3, the norm of projector $\boldsymbol{Q}$ is proportional to the ratio β. Our approach defines the modified direction $\boldsymbol{R}$ of weights as:

$$\boldsymbol{R} = \boldsymbol{P}_{norm} + \mathrm{m}\boldsymbol{Q}_{norm} \tag{5}$$

$$\boldsymbol{P}_{norm} = \frac{\boldsymbol{P}}{||\boldsymbol{P}||}, \quad \boldsymbol{Q}_{norm} = \frac{\boldsymbol{Q}}{||\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}||} \tag{6}$$

where m is a constant to constrain the norm of projector $\boldsymbol{Q}$ to prevent gradient explosion or gradient vanishing in the backward process; $\boldsymbol{P}_{norm}$ and $\boldsymbol{Q}_{norm}$ are identity matrices normalized by $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively. Normalization is to prevent the case that the change of β has little effect on the modified direction because of the large norm gap between $\boldsymbol{P}$ and $\boldsymbol{Q}$. In the back-propagate (BP) process, the direction of network weights is modified as:

$$\begin{aligned} \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \boldsymbol{G} \quad j = 1 \\ \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \boldsymbol{R}_l(j-1)\boldsymbol{G} \quad j > 1 \\ \boldsymbol{G} &= \gamma(i,j)\Delta\boldsymbol{W}_l^{BP}(i,j) \end{aligned} \tag{7}$$

where $\boldsymbol{W}_l(i,j) \in \mathbb{R}^{s\times v}$ represents the connection weights between the lth layer and the $(l+1)$th layer; γ represents the learning rate of this network; $\Delta\boldsymbol{W}_l^{BP}(i,j)$ represents the standard BP gradient; $\boldsymbol{R}$ represents the modification projector in our method. In Eq 7, we can easily observe that we modify weight direction adaptively by multiplying the BP gradient $\Delta\boldsymbol{W}_l^{BP}(i,j)$ with our projector $\boldsymbol{R}$ whose direction is varied according to the ratio β of classes with similar feature distribution between different datasets and others.

Algorithm 1 Regularized Adaptive Weight Modification

Require: Training data from different datasets, γ (learning rate), m (constant hyperparameter), T_{reg} (constant hyperparameter).
for every dataset *j* **do**
  for every batch *i* **do**
    if $j = 1$ **then**
      $\boldsymbol{W}_l(i,j) = \boldsymbol{W}_l(i-1,j) + \gamma(i,j)\Delta\boldsymbol{W}_l^{BP}(i,j)$
    else
      $\mathbf{k}(i,j) = \frac{\boldsymbol{P}_l(i-1)\overline{\mathbf{x}}_{l-1}(i,j)}{\alpha + \overline{\mathbf{x}}_{l-1}(i,j)^T\boldsymbol{P}_l(i-1,j)\overline{\mathbf{x}}_{l-1}(i,j)}$
      $\boldsymbol{P}_l(i,j) = \boldsymbol{P}_l(i-1,j) - \mathbf{k}(i,j)\overline{\mathbf{x}}_{l-1}(i,j)^T\boldsymbol{P}_l(i-1,j)$
      $\beta = \frac{\sum_{t=1}^{b} N_t + 1}{\sum_{t=b+1}^{b+c} N_t + 1}$
      $\boldsymbol{Q} = \beta[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}]$
      $\boldsymbol{P}_N = \frac{\boldsymbol{P}}{||\boldsymbol{P}||}$
      $\boldsymbol{Q}_N = \frac{\boldsymbol{Q}}{||\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}||}$
      $\boldsymbol{R} = \boldsymbol{P}_N + \mathrm{m}\boldsymbol{Q}_N$
      $\hat{y}_o(i) = \frac{y_o(i)^{1/T_{reg}}}{\sum \mathbf{y}_o(i)^{1/T_{reg}}}$
      $\hat{y}_n(i) = \frac{y_n(i)^{1/T_{reg}}}{\sum \mathbf{y}_n(i)^{1/T_{reg}}}$
      $\Delta\boldsymbol{W}_{l_{reg}}^{BP} = -\nabla(\hat{\mathbf{y}}_o(i) \cdot \log \hat{\mathbf{y}}_n(i))$
      $\boldsymbol{G} = \gamma(i,j)\Delta\boldsymbol{W}_l^{BP}(i,j)$
      $\boldsymbol{H} = (1-\eta)\boldsymbol{R}_l(j-1)\boldsymbol{G} + \eta\Delta\boldsymbol{W}_{l_{reg}}^{BP}(i,j)$
      $\boldsymbol{W}_l(i,j) = \boldsymbol{W}_l(i-1,j) + \boldsymbol{H}$
    end if
  end for
end for

4.2. Regularization

There are also a few datasets where genuine audio is collected from quite different acoustic conditions compared with others. In this case, it is unreasonable to use the above method directly. As for these utterances, we introduce an extra regularization forcing the model to remember the previous inference distribution.

We first replicate the pre-trained model into two models

with the same parameters, one is the teacher model and the other one is the student model. The parameter of the teacher model is frozen in the process of training on the new dataset and the parameter of the student model is fine-tuned. Like the operation in the LwF, we view the softmax output $\mathbf{y}_o$ from the teacher model as "soft labels" and use the loss function to slash the distinction between the "soft labels" $\mathbf{y}_o$ and the softmax output $\mathbf{y}_n$ of the student model, thus forcing the student model to remember the learned knowledge. The loss function, which is a modified cross-entropy loss, can be written as:

$$\boldsymbol{L}_{reg}(\hat{\mathbf{y}}_o, \hat{\mathbf{y}}_n) = -\hat{\mathbf{y}}_o \cdot \log \hat{\mathbf{y}}_n \tag{8}$$

$$\hat{y}_o = \frac{y_o^{1/T_{reg}}}{\sum \mathbf{y}_o^{1/T_{reg}}}, \quad \hat{y}_n = \frac{y_n^{1/T_{reg}}}{\sum \mathbf{y}_n^{1/T_{reg}}} \tag{9}$$

where T_{reg} is a constant hyperparameter. The $\mathbf{y_o}$, $\mathbf{y_n}$ are softmax outputs of teacher and student models, respectively; The $\hat{\mathbf{y}}$ is a normalized form of the $\mathbf{y}$; The $\hat{y}$ and y are one item of $\hat{\mathbf{y}}$ and $\mathbf{y}$, respectively. The weight modification of this regularization $\Delta \boldsymbol{W}_{l_{reg}}^{BP}$ can be written as Eq 10.

$$\Delta \boldsymbol{W}_{l_{reg}}^{BP} = \nabla \boldsymbol{L}_{reg} \tag{10}$$

4.3. Regularized Adaptive Weight Modification

In brief, our method RAWM is proposed for general continual learning conditions by modifying weight direction according to the ratio β of classes with similar feature distribution between different datasets and others in batch data. By introducing a regularized restriction, our method eases the problem that a few data belonging to classes in the numerator of the Eq 4 may have distinct feature distributions because they are collected from quite different conditions. Our method is inspired by fake audio detection where the ratio β in the Eq 4 can be written as:

$$\beta = \frac{N_g + 1}{N_f + 1} \tag{11}$$

in which N_g and N_f represent the number of genuine and fake audios in a batch, respectively. As for another research area, for example, speech emotion recognition including happy, sad, angry, and neutral, the essential ratio can be written as:

$$\beta = \frac{N_{neu} + 1}{N_{ang} + N_{hap} + N_{sad} + 1} \tag{12}$$

because the neutral emotion has a relatively more similar feature distribution than others between different datasets. The N_{neu}, N_{ang}, N_{sad}, and N_{hap} represent the number of neutral, angry, sad, and happy data in a batch, respectively.

Considering a continual learning situation, the BP process of regularized adaptive weight modification can be written as Eq 13.

$$\begin{aligned} \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \boldsymbol{G} \qquad j = 1 \\ \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \boldsymbol{H} \qquad j > 1 \\ \boldsymbol{G} &= \gamma(i,j) \Delta \boldsymbol{W}_l^{BP}(i,j) \\ \boldsymbol{H} &= (1-\eta)\boldsymbol{R}_l(j-1)\boldsymbol{G} + \eta \Delta \boldsymbol{W}_{l_{reg}}^{BP}(i,j) \end{aligned} \tag{13}$$

Compared with the Eq 7, our method introduces a regularization constraint to the adaptive weight modification. The importance of the regularization depends on the hyperparameter η which is a coefficient measuring the attention degree of the knowledge acquired from old datasets.

5. Experiments

5.1. Datasets

We conduct our experiments on four fake audio datasets, including the ASVspoof2019LA ($\mathbf{S}$), ASVspoof2015 ($\mathbf{T_1}$), VCC2020 ($\mathbf{T_2}$), and In-the-Wild ($\mathbf{T_3}$). The models are firstly trained using the training set of the $\mathbf{S}$ and are fine-tuned on the training sets of the other three datasets. All of the experiments are evaluated using two or four evaluation sets in these datasets. The final model in the study refers to the model that was trained after the entire training process and then evaluated on each dataset.

ASVspoof 2019 LA Dataset (Todisco et al., 2019) is the sub-challenge dataset (30 males and 37 females) containing three subsets: training, development, and evaluation. The training set and development share the same attack including four TTS and two VC algorithms. The bonafide audio is collected from the VCTK corpus (Veaux et al., 2017). The evaluation set contains totally different attacks.

ASVspoof2015 dataset (Wu et al., 2015) is an open-source standard dataset of genuine and synthetic speech in the ASVspoof2015 challenge. The genuine speech was recorded from 106 speakers (45 males and 61 females) with no significant channel or background noise effects. The spoofing speech is generated using a variety of speech synthesis and voice conversion algorithms.

VCC2020 dataset (Zhao et al., 2020) is collected from Voice Conversion Challenge 2020. This dataset contains two subsets: a set of genuine audio provided by organizers and a set of fake audio provided by participating teams. Different from the previous three datasets, VCC2020 is a multilingual fake audio dataset, including English, Finnish, German and Mandarin.

In-the-Wild dataset (Müller et al., 2022) contains a set of deep fake audio (and corresponding real audio) of 58 politicians and other public figures collected from publicly available sources, such as social networks and video streaming platforms. In total, 20.8 hours of genuine audio and 17.2 hours of fake audio were collected. On average, each speaker had 23 minutes of genuine audio and 18 minutes of fake audio.

We divide the genuine and fake audios of the VCC2020 dataset into four subsets. A quarter is used to build the evaluation set, a quarter to build the development set, and the rest to be used as the training set. The In-the-Wild dataset

Table 1: (a) is the statistics of experimental datasets and (b) is the EER(%) of our baseline on multiple evaluation sets.

(a)

Dataset	ASVSpoof2019LA (**S**)		ASVSpoof2015 ($\mathbf{T_1}$)		VCC2020 ($\mathbf{T_2}$)		In-the-Wild ($\mathbf{T_3}$)	
	#Real	#Fake	#Real	#Fake	#Real	#Fake	#Real	#Fake
Train	2,580	22,800	3,750	12,625	1,330	3,060	9,431	5,908
Dev	2,548	22,296	3,497	49,875	665	1,530	4,715	2,954
Eval	7,355	63,882	9,404	184,000	665	1,530	4,717	2,954

(b)

Model	**S**	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Baseline	0.258	24.532	46.503	91.473

Table 2: The EER(%) on evaluation sets of our method with different η. (a), (b) and (c) are trained using the training set in order to $\mathbf{S} \rightarrow \mathbf{T_k}$ and are evaluated using the evaluation set on **S** and $\mathbf{T_k}$; (d) is trained using training set in order to $\mathbf{S} \rightarrow \mathbf{T_1} \rightarrow \mathbf{T_2} \rightarrow \mathbf{T_3}$ and is evaluated using evaluation sets.

(a)

η	**S**	$\mathbf{T_1}$
Baseline	0.258	24.532
0.00	1.643	0.256
0.20	1.424	0.431
0.25	1.175	0.311
0.50	0.878	0.257
0.75	**0.666**	**0.247**
1.00	3.123	0.343

(b)

η	**S**	$\mathbf{T_2}$
Baseline	0.258	46.503
0.00	1.413	3.845
0.20	1.334	4.288
0.25	1.275	3.994
0.50	**1.237**	**3.721**
0.75	1.262	4.571
1.00	4.234	4.566

(c)

η	**S**	$\mathbf{T_3}$
Baseline	0.258	91.473
0.00	6.126	3.457
0.20	5.490	3.848
0.25	4.975	3.593
0.50	**4.942**	**3.249**
0.75	4.482	4.271
1.00	4.453	4.598

(d)

η	**S**	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Baseline	0.258	24.532	46.503	91.473
0.00	1.845	1.127	3.916	3.410
0.20	1.724	1.003	4.120	3.367
0.25	1.699	0.945	4.017	3.529
0.50	**1.508**	**0.641**	**3.850**	**3.163**
0.75	1.636	0.873	3.975	4.454
1.00	2.714	1.621	3.875	4.325

is divided in the same way as the VCC2020. The detailed statistics of the datasets are presented in Table 1a. The Equal Error Rate (EER), which is widely used for fake audio detection and speaker verification, is applied to evaluate the experimental performance.

5.2. Experimental Setup

Fake Audio Detection Model: We use the pre-trained model Wav2vec 2.0 (Baevski et al., 2020) as the feature extractor and the self-attention convolutional neural network (S-CNN) as the classifier. The parameters of Wav2vec 2.0 is loaded from the pre-train model XLSR-53 (Conneau et al., 2020). The classifier S-CNN contains three 1D-Convolution layers, one self-attention layer, and two full connection layers, according to the forward process. The input dimension of the first convolution layer is 256 and the hidden dimension of all convolution layers is 80. The kernel size and stride are set to 5 and 1, respectively. The hidden dimension of all full connection layers is 80 and the output dimension is 2.

Training Details: We fine-tune the model weights including the pre-trained model XLSR-53 and the classifier S-CNN. All of the parameters are trained by the Adam optimizer with a batch size of 2 and a learning rate γ of 0.0001. The constant m and T_{reg} in RAWM are set to 0.1 and 2, respectively. The α is initialized to 0.00001 for convolution layers, 0.0001 for the self-attention layer, and 0.1 for full connection layers. The norm in normalization of projector $\boldsymbol{P}$ and $\boldsymbol{Q}$ is the $\boldsymbol{L}^2$ norm. In addition, we present the results of training all datasets (Tain-on-All) that is considered to be the lower bound to all continual learning methods we mentioned (Parisi et al., 2019). All results are (re)produced by us and averaged over 7 runs with standard deviations.

5.3. Baseline

We first train our model on the training set of the **S** dataset. Table 1b shows the detection performance of our baseline on multiple evaluation sets which is very close to the state-of-the-art result (Nautsch et al., 2021) in the same dataset. Although the model achieves promising performance on **S**, its detection accuracy degrades significantly on other datasets. In addition, our baseline achieves the lowest cross-datasets EER on $\mathbf{T}_1$ dataset among three unseen datasets, which verifies that the detection model will have better performance when facing genuine audio with more similar feature distribution. Apart from that, the results with different training steps are presented in Table 8 in the appendix.

5.4. Effects of the η for our method

Sequence training between two datasets: We start by performing some experiments to evaluate the effectiveness of η in RAWM, which represents the attention degree to learned knowledge. In Table 2, we can easily observe that the RAWM achieves great performance on both old and new datasets, especially in the experiment on $\mathbf{S} \rightarrow \mathbf{T}_1$. By comparing the results of three cross-datasets, we observe that when the new and old datasets have similar feature distribu-

Table 3: The EER(%) on evaluation sets of the ablation studies. (a), (b) and (c) are trained using the training set in order to $\mathbf{S} \rightarrow \mathbf{T_k}$ and are evaluated using the evaluation set on $\mathbf{S}$ and $\mathbf{T_k}$; (d) is trained in order to $\mathbf{S} \rightarrow \mathbf{T_1} \rightarrow \mathbf{T_2} \rightarrow \mathbf{T_3}$ and are evaluated using evaluation sets.

(a)

Method	S	$\mathbf{T_1}$
RAWM	**0.666**	**0.247**
–REG	1.643	0.256
–AWM	2.448	0.500

(b)

Method	S	$\mathbf{T_2}$
RAWM	**1.237**	**3.721**
–REG	1.413	3.845
–AWM	3.086	5.432

(c)

Method	S	$\mathbf{T_3}$
RAWM	**4.942**	**3.249**
–REG	7.126	3.357
–AWM	8.130	5.065

(d)

Method	S	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
RAWM	**1.508**	**0.641**	**3.850**	**3.163**
–REG	1.845	1.127	3.916	3.410
–AWM	4.083	2.167	6.480	5.472

Table 4: The EER(%) of our method compared with various methods. (a), (b) and (c) are trained using the training set in order to $\mathbf{S} \rightarrow \mathbf{T_k}$ and are evaluated using the evaluation set on $\mathbf{S}$ and $\mathbf{T_k}$; (d) is trained using training set in order to $\mathbf{S} \rightarrow \mathbf{T_1} \rightarrow \mathbf{T_2} \rightarrow \mathbf{T_3}$ and is evaluated using evaluation sets.

(a)

Method	S	$\mathbf{T_1}$
Baseline	0.258	24.532
Train-on-All	0.406	0.201
Fine-tune	7.324	0.510
EWC	2.832	0.570
OWM	2.448	0.540
LwF	3.123	0.343
DFWF	1.849	0.689
RAWM(Ours)	**0.666**	**0.247**

(b)

Method	S	$\mathbf{T_2}$
Baseline	0.258	46.503
Train-on-All	0.965	2.498
Fine-tune	8.755	5.647
EWC	3.494	6.289
OWM	3.086	6.432
LwF	4.234	4.566
DFWF	1.874	7.355
RAWM(Ours)	**1.237**	**3.721**

(c)

Method	S	$\mathbf{T_3}$
Baseline	0.258	91.473
Train-on-All	2.740	2.160
Fine-tune	20.976	4.978
EWC	8.039	5.615
OWM	8.130	5.065
LwF	6.453	4.998
DFWF	4.324	6.275
RAWM(Ours)	**4.942**	**3.249**

(d)

Method	S	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Baseline	0.258	24.532	46.503	91.473
Train-on-All	1.324	0.561	3.579	2.008
Fine-tune	7.068	2.841	5.674	4.543
EWC	5.569	2.444	6.510	5.129
OWM	4.083	2.167	6.480	5.472
LwF	2.714	1.621	4.875	4.325
DFWF	3.476	3.735	7.345	6.114
RAWM(Ours)	**1.508**	**0.641**	**3.850**	**3.163**

tion (Table 2a), there is an improvement in the performance of both acquiring new knowledge and overcoming forgetting with the increasing of η ($\eta < 1$); When the feature distribution of the new and old datasets is different (Table 2b, Table 2c), it is the model when $\eta = 0.50$ that achieves the best result, which shows that the regularization we introduced is also of benefit to performance on both learning and overcoming forgetting.

Sequence training on four datasets: We also present the results on multiple evaluation sets about different η in Table 2d. It can be observed that our method slashes performance degradation when training across datasets. The RAWM achieves the lowest EER among the results when $\eta = 0.50$, which demonstrates that the same attention degree to both old and new datasets is the best choice for learning and overcoming forgetting. In addition, the results of $\mathbf{S}$, $\mathbf{T_1}$ and $\mathbf{T_2}$ show that the model with larger η is more effective in overcoming forgetting.

5.5. Ablation studies for our method

Sequence training between two datasets: In this section, we compare our proposed method with adaptive weight modification without regularization (–REG) and orthogonal weight modification without regularization (–AWM). Table 3 presents their EER on three evaluation sets. We observe that RAWM achieves similar EER to –REG on the new dataset, both of them are superior significantly to –AWM, which shows that the adaptive weight modification has a significant positive impact on acquiring knowledge, while regularization impacts little. As for overcoming forgetting, when the feature distribution of the new and old datasets is similar (Table 3a), the EER of the –REG on the old datasets is much lower than that of the –AWM and higher than that of the RAWM, which shows that the adaptive weight modification and regularization can significantly reduce the forgetting in this case. When the languages of the new and old datasets are different (Table 3b), the EER of RAWM in the old datasets is similar to that of the –REG and much lower than that of the –AWM, which also proves that the adaptive weight modification has a significant positive impact on overcoming forgetting. When the feature distribution of the new and old datasets is quite different (Table 3c), the EER of the –REG is similar to that of the –AWM and much higher than that of the RAWM, which shows that in this case, regularization is of great benefit to overcoming forgetting, while the effect of adaptive weight modification is not obvious.

Sequence training on four datasets: In this section, we present the results of the ablation study on four evaluation sets in Table 3d. We observe that the EER of –REG to –AWM degrades more obviously than that of RAWM to –REG on all evaluation sets, which indicates that adaptive weight modification has a more obvious benefit in learning and overcoming forgetting than regularization for sequence training on multiple datasets.

Table 5: The EER(%) of few samples experiments. All experiments are first trained using the training set of $\mathbf{S}$ and then trained on 100 samples of the training set of $\mathbf{T}_1$. All experiments are evaluated using the evaluation set on $\mathbf{S}$ and $\mathbf{T}_1$.

Method	$\mathbf{S}$	$\mathbf{T}_1$
Baseline	0.258	24.532
Train-on-All	0.279	0.291
Fine-tune	7.951	0.617
EWC	2.972	0.619
OWM	2.683	0.617
LwF	3.198	0.542
DFWF	1.975	0.733
RAWM(Ours)	**0.923**	**0.312**

Table 6: The Acc(%) of various continual learning methods for 4-classes speech emotion recognition. All experiments are trained using the training set in order to $\mathbf{MSP\text{-}Podcast} \rightarrow \mathbf{IEMOCAP}$ and are evaluated using the evaluation set on $\mathbf{MSP\text{-}Podcast}$ and $\mathbf{IEMOCAP}$

Method	MSP-Podcast	IEMOCAP
Baseline	54.446	30.043
Train-on-All	54.396	57.262
Fine-tune	24.094	50.379
EWC	35.819	48.698
OWM	32.267	48.162
LwF	38.800	44.034
RAWM(Ours)	**41.995**	**54.229**

5.6. Comparison of our method with other methods

Sequence training between two datasets: We compare our method with several methods in Table 4. The EWC, LwF, and OWM as three mainstream continual learning methods achieve great success in many fields. The DFWF is the first continual learning method to overcome forgetting for fake audio detection. The results demonstrate that fine-tuning without modification (Fine-tune) forgets previous knowledge obviously. The forgetting of RAWM is one-tenth that of Fine-tune on Table 4a and the EER on the new dataset of RAWM is also half that of Fine-tune. We also observe that the Fine-tune, EWC and OWM achieve similar performance in three experiments and the performance of LwF outperforms theirs on the new dataset. The DFWF is more effective in overcoming forgetting than the above methods, but its performance on the new dataset is inferior to others. Compared with others, our method achieves lower EER on both old and new datasets of all experiments, which demonstrates that both overcoming forgetting and learning could definitely benefit from our method when training across datasets, regardless of whether the datasets have similar feature distributions (Table 4a, Table 4b) or same languages (Table 4c).

Sequence training on four datasets: In addition, We compare our method with several methods for sequence training on four datasets in Table 4d. The results show that most methods achieve lower EERs than fine-tuning, and the best result for overcoming forgetting and learning is our proposed method, which indicates that the RAWM is superior to others for sequence training on both two and multiple datasets.

5.7. The performance of the RAWM with a few samples

We also present some results of the model when training on a few samples of new datasets. In our experiments, only 100 samples randomly selected from new datasets $\mathbf{T}_1$ were used for fine-tuning or continual learning. All models are first trained on the $\mathbf{S}$ datasets and then fine-tuned or continually learned on the $\mathbf{T}_1$ dataset. All models are trained on the new dataset within five steps. From the results, we can observe that our method RAWM also achieves the best performance on both old and new datasets and the learning performance is very close to the result of Train-on-All which is the the lower bound to all continual learning methods. By comparing the results in Table 4a and Table 5, we can easily find that reducing the number of samples has only a little damage to our method.

5.8. The RAWM for speech emotion recognition

Our method is inspired by fake audio detection and it can be easily used in other related tasks. We take speech emotion recognition as an example to evaluate the performance of the RAWM in other fields. In this regard, the previous result shows that neutral emotion achieved the highest recognition accuracy across thirteen emotion datasets (Sharma, 2022). So we infer that neutral speech has a more similar feature distribution than that of happy, sad, and angry, thus the ratio β of our method can be written as Eq 12. Based on this observation, we conduct some experiments for speech emotion recognition. We choose four emotional classes, including neutral, happy, angry, and sad. The feature extractor and classifier are as same as that in fake audio detection. The results have been shown in Table 6. It could be easily observed that our method still achieves the highest accuracy on both datasets. The effect of our method in overcoming forgetting is most obvious and its learning performance is very close to the result of Train-on-All.

5.9. The RAWM for image recognition

We also have performed evaluation experiments on the image recognition domain in the CLEAR benchmark (Lin et al., 2021) to explore the broader applicability of our method. The CLEAR-10 benchmark for continual learning consists of 10 image recognition experiences, each comprising 11

Table 7: The Accuracy(%) on the CLEAR experiences.

Continual Learning Methods	Acc on the evaluation set of each experience									
	$\mathbf{Exp_1}$	$\mathbf{Exp_2}$	$\mathbf{Exp_3}$	$\mathbf{Exp_4}$	$\mathbf{Exp_5}$	$\mathbf{Exp_6}$	$\mathbf{Exp_7}$	$\mathbf{Exp_8}$	$\mathbf{Exp_9}$	$\mathbf{Exp_{10}}$
Replay	**94.34**	**93.64**	**94.34**	**95.15**	**94.75**	**94.55**	**94.34**	**94.34**	**95.35**	**96.06**
Fine-tune	87.68	90.00	91.11	91.82	90.40	89.90	90.30	90.61	90.61	93.33
EWC	84.04	84.95	85.86	87.07	85.66	85.56	86.97	86.16	85.76	87.78
LwF	88.59	88.89	87.27	90.51	87.68	87.78	87.47	87.47	88.79	88.48
GDF	91.11	91.62	88.38	91.01	88.79	89.19	90.20	87.68	90.10	90.30
CWR	90.71	91.72	90.71	91.52	89.49	90.91	91.62	90.71	91.82	93.74
OWM	91.62	92.12	**91.82**	93.64	91.72	**92.42**	92.22	**92.32**	92.42	95.05
RAWM (Ours)	**92.12**	**92.53**	91.41	**93.74**	**91.82**	**92.42**	**92.53**	92.22	**92.53**	**95.25**

classes such as camera, baseball, laptop, etc. To evaluate the effectiveness of our method, we selected several widely used continual learning algorithms, including the Replay, EWC, LwF, GDumbFinetune (GDF) (Prabhu et al., 2020), CopyWeights with Re-init (CWR) (Lomonaco & Maltoni, 2017), and OWM methods. Table 7 presents the results of the comparative analysis. We treated the Replay method, which corresponds to the "Train-on-all" approach in our paper, as the upper bound of accuracy for all continual learning methods. The EWC and LwF methods have been introduced in our paper. In the GDF algorithm, we set the memory size to be the same as the number of training data in one bucket, and for CWR, the cwr layer was positioned as the final layer of the model. To extract features, we employed a pre-trained ResNet-50 (He et al., 2016) as a feature extractor, producing 2048-dimensional feature vectors. A linear layer with input and output dimensions of 2048 and 11, respectively, was used as the downstream classifier. Our experiments were conducted with a batch size of 512, an initial learning rate of 1 (decayed by a factor of 0.1 after 60 epochs), and the SGD optimizer with a momentum of 0.9. The experimental results in Table 7 demonstrate that our proposed method, referred to as RAWM, consistently achieved the best performance across most tasks. In particular, in $\mathbf{Exp_3}$ and $\mathbf{Exp_8}$, the performance of our method closely approached the highest accuracy achieved among all the evaluated methods.

6. Conclusion

In this work, we propose a continual learning algorithm to overcome catastrophic forgetting, called RAWM, that could adaptively modify the weight direction in process of training on new datasets. We also introduce a regularization to deal with the situation when old and new datasets are collected from quite different conditions. The experimental results demonstrate that our method outperforms four continual learning methods in learning and overcoming forgetting in scenarios of sequence training on both two and multiple datasets. The result shows that our method still achieves the best performance among the above methods when training on a few samples. Besides, our method is inspired by fake audio detection and the results show that it can be easily generalized to other fields, like speech emotion recognition. In addition, our method does not require previous data; thus it can be applied to most classification networks. We have yet to study how to make the model learn the weight direction gradually in the process of training on new datasets without any constraint, and exploring generalization to related tasks will form the focus of our future studies.

Acknowledgements

This work is supported by the National Key Research and Development Program of China under Grant No.2020AAA0140003, the National Natural Science Foundation of China (NSFC) (No.61831022, No.U21B2010, No.62101553, No.61971419, No.62006223, No.62276259, No.62201572, No.62206278), Beijing Municipal Science and Technology Commission, Administrative Commission of Zhongguancun Science Park No.Z211100004821013, Open Research Projects of Zhejiang Lab (NO.2021KH0AB06).

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Aljundi, R., Kelchtermans, K., and Tuytelaars, T. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11254–11263, 2019.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Ben-Israel, A. and Greville, T. N. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

Bengio, Y. and LeCun, Y. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., Schmid, C., and Alahari, K. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Cermelli, F., Mancini, M., Bulo, S. R., Ricci, E., and Caputo, B. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9233–9242, 2020.

Chen, Y., Diethe, T., and Lawrence, N. Facilitating bayesian continual learning by natural gradients and stein gradients. *arXiv preprint arXiv:1904.10644*, 2019.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

de Masson D'Autume, C., Ruder, S., Kong, L., and Yogatama, D. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Haykin, S. S. *Adaptive filter theory*. Pearson Education India, 2002.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huszár, F. On quadratic penalties in elastic weight consolidation. *CoRR*, abs/1712.03847, 2017. URL http://arxiv.org/abs/1712.03847.

Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N. W. D., Yamagishi, J., and Lee, K. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Lacerda, F. (ed.), *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pp. 2–6. ISCA, 2017. URL http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1111.html.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, K., Lee, K., Shin, J., and Lee, H. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 312–321, 2019.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Lin, Z., Shi, J., Pathak, D., and Ramanan, D. The clear benchmark: Continual learning on real-world imagery. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Liu, T., Ungar, L., and Sedoc, J. Continual learning for sentence representations using conceptors. *arXiv preprint arXiv:1904.09187*, 2019.

Lomonaco, V. and Maltoni, D. Core50: a new dataset and benchmark for continuous object recognition. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pp. 17–26. PMLR, 2017. URL http://proceedings.mlr.press/v78/lomonaco17a.html.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Lv, Z., Zhang, S., Tang, K., and Hu, P. Fake audio detection based on unsupervised pretraining models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9231–9235. IEEE, 2022.

Ma, H., Yi, J., Tao, J., Bai, Y., Tian, Z., and Wang, C. Continual learning for fake audio detection. *arXiv preprint arXiv:2104.07286*, 2021.

Ma, H., Yi, J., Wang, C., Yan, X., Tao, J., Wang, T., Wang, S., Xu, L., and Fu, R. FAD: A chinese dataset for fake audio detection. *CoRR*, abs/2207.12308, 2022. doi: 10.48550/arXiv.2207.12308. URL https://doi.org/10.48550/arXiv.2207.12308.

Mallya, A. and Lazebnik, S. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Martín-Doñas, J. M. and Álvarez, A. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9241–9245. IEEE, 2022.

Monteiro, J., Alam, M. J., and Falk, T. H. An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 6599–6603. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9054558. URL https://doi.org/10.1109/ICASSP40776.2020.9054558.

Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., and Böttinger, K. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Nautsch, A., Wang, X., Evans, N. W. D., Kinnunen, T. H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., and Lee, K. A. Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(2):252–265, 2021. doi: 10.1109/TBIOM.2021.3059479. URL https://doi.org/10.1109/TBIOM.2021.3059479.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi: 10.1016/j.neunet.2019.01.012. URL https://doi.org/10.1016/j.neunet.2019.01.012.

Perez-Rua, J.-M., Zhu, X., Hospedales, T. M., and Xiang, T. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13846–13855, 2020.

Prabhu, A., Torr, P. H. S., and Dokania, P. K. Gdumb: A simple approach that questions our progress in continual learning. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J. (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pp. 524–540. Springer, 2020. doi: 10.1007/978-3-030-58536-5_31. URL https://doi.org/10.1007/978-3-030-58536-5_31.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537. PMLR, 2018.

Serra, J., Suris, D., Miron, M., and Karatzoglou, A. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pp. 4548–4557. PMLR, 2018.

Shah, S., Palmieri, F., and Datum, M. Optimal filtering algorithms for fast learning in feedforward neural networks. *Neural networks*, 5(5):779–787, 1992.

Sharma, M. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6907–6911, 2022. doi: 10.1109/ICASSP43922.2022.9747417.

Tak, H., Todisco, M., Wang, X., Jung, J.-w., Yamagishi, J., and Evans, N. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., and Lee, K. A. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

Tseran, H., Khan, M. E., Harada, T., and Bui, T. D. Natural variational continual learning. In *Continual Learning Workshop@ NeurIPS*, volume 2, 2018.

Veaux, C., Yamagishi, J., MacDonald, K., et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

Wang, H., Dinkel, H., Wang, S., Qian, Y., and Yu, K. Dual-adversarial domain adaptation for generalized replay attack detection. In Meng, H., Xu, B., and Zheng, T. F. (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association,*

Virtual Event, Shanghai, China, 25-29 October 2020, pp. 1086–1090. ISCA, 2020. doi: 10.21437/Interspeech.2020-1255. URL https://doi.org/10.21437/Interspeech.2020-1255.

Wang, T., Fu, R., Yi, J., Tao, J., Wen, Z., Qiang, C., and Wang, S. Prosody and voice factorization for few-shot speaker adaptation in the challenge m2voc 2021. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8603–8607, 2021.

Wang, X. and Yamagishi, J. Investigating self-supervised front ends for speech spoofing countermeasures. *arXiv preprint arXiv:2111.07725*, 2021.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, 2018.

Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., and Fu, Y. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019.

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., and Sizov, A. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth annual conference of the international speech communication association*, 2015.

Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

Yan, X., Yi, J., Tao, J., Wang, C., Ma, H., Wang, T., Wang, S., and Fu, R. An initial investigation for detecting vocoder fingerprints of fake audio. *CoRR*, abs/2208.09646, 2022. doi: 10.48550/arXiv.2208.09646. URL https://doi.org/10.48550/arXiv.2208.09646.

Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., Wang, T., Tian, Z., Bai, Y., Fan, C., et al. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9216–9220. IEEE, 2022.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

Zeng, G., Chen, Y., Cui, B., and Yu, S. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017a. URL http://proceedings.mlr.press/v70/zenke17a.html.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017b.

Zhang, Y., Jiang, F., and Duan, Z. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021a.

Zhang, Y., Zhu, G., Jiang, F., and Duan, Z. An empirical study on channel effects for synthetic voice spoofing countermeasure systems. *arXiv preprint arXiv:2104.01320*, 2021b.

Zhao, Y., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., and Toda, T. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*, 2020.

A. Appendix

Table 8: The EER(%) on multiple evaluation sets. Model-1 to Model-6 are the models trained using the ASVspoof2019LA training set with increasing training steps.

Model	Evaluation Sets			
	S	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Model-1	3.751	**6.316**	**7.670**	**75.198**
Model-2	2.975	8.517	10.000	78.477
Model-3	1.794	9.988	26.165	85.436
Model-4	**0.258**	24.532	46.503	91.473
Model-5	0.259	25.698	44.741	91.824
Model-6	0.262	27.872	49.726	92.113