Region-Based Optimization in Continual Learning for Audio Deepfake Detection

Yujie Chen [1], Jiangyan Yi [3,*], Cunhang Fan [1], Jianhua Tao [3,4], Yong Ren [2], Siding Zeng [2], Chu Yuan Zhang [3], Xinrui Yan [2], Hao Gu [2], Jun Xue [1], Chenglong Wang [2], Zhao Lv [1], Xiaohui Zhang [2,*]

[1] School of Computer Science and Technology, Anhui University, China
[2] Institute of Automation, Chinese Academy of Sciences
[3] Department of Automation, Tsinghua University
[4] Beijing National Research Center for lnformation Science and Technology, Tsinghua University
e22201148@stu.ahu.edu.cn, yijy@tsinghua.edu.cn

Abstract

Rapid advancements in speech synthesis and voice conversion bring convenience but also new security risks, creating an urgent need for effective audio deepfake detection. Although current models perform well, their effectiveness diminishes when confronted with the diverse and evolving nature of real-world deepfakes. To address this issue, we propose a continual learning method named Region-Based Optimization (RegO) for audio deepfake detection. Specifically, we use the Fisher information matrix to measure important neuron regions for real and fake audio detection, dividing them into four regions. First, we directly fine-tune the less important regions to quickly adapt to new tasks. Next, we apply gradient optimization in parallel for regions important only to real audio detection, and in orthogonal directions for regions important only to fake audio detection. For regions that are important to both, we use sample proportion-based adaptive gradient optimization. This region-adaptive optimization ensures an appropriate trade-off between memory stability and learning plasticity. Additionally, to address the increase of redundant neurons from old tasks, we further introduce the Ebbinghaus forgetting mechanism to release them, thereby promoting the model's ability to learn more generalized discriminative features. Experimental results show our method achieves a 21.3 percent improvement in EER over the state-of-the-art continual learning approach RWM for audio deepfake detection. Moreover, the effectiveness of RegO extends beyond the audio deepfake detection domain, showing potential significance in other tasks, such as image recognition.

Introduction

Recently, with the rapid development of speech synthesis and voice conversion technologies, the distinction between real and fake audio has become increasingly blurred, posing significant security risks to society. Consequently, researchers are increasingly focusing on audio deepfake detection mechanisms (Yi et al. 2023b). Community-led initiatives, such as the ASVspoof Challenges (Kinnunen et al. 2017; Todisco et al. 2019; Yamagishi et al. 2021) and the Audio Deepfake Detection (ADD) Challenges (Yi et al. 2022, 2023a), significantly advance the field of fake audio detection. In addition, the introduction of pre-trained audio models significantly improves the effectiveness of audio deepfake detection, achieving impressive performance on publicly available datasets. (Tak et al. 2022; Wang and Yamagishi 2022; Hsu et al. 2021)

Despite the significant advancements in audio deepfake detection models, their performance is still limited when confronted with diverse and unseen forged audio in real-world scenarios. To address this challenge, two primary approaches have been developed. The first approach utilizes data augmentation and multi-feature fusion techniques to extract robust audio features, improving the generalization of model across various datasets (Wang et al. 2023; Fan et al. 2024). The second approach is based on continual learning (Ma et al. 2021), where the model incrementally learns from both new and old datasets, allowing it to integrate previously learned discriminative information. This enhances its detection capability for diverse and unseen deepfakes, achieving a balance between **memory stability** (the model's ability to retain performance on old tasks) and **learning plasticity** (the model's ability to perform on new tasks). Currently, the most advanced continual learning methods for audio deepfake detection, Regularized Adaptive Weight Modification (RAWM) (Zhang et al. 2023) and Radian Weight Modification (RWM) (Zhang et al. 2024), overcome catastrophic forgetting by introducing trainable gradient correction directions to optimize weights. While RAWM and RWM exhibit notable effectiveness in overcoming catastrophic forgetting, the use of Recursive Least Squares (RLS) (Shah, Palmieri, and Datum 1992) to approximate the gradient plane of previous tasks can introduce errors (Liavas and Regalia 1999). Moreover, applying gradient modification to all neurons restricts the model's learning plasticity for new tasks.

To address the aforementioned issues, we propose a continual learning method for audio deepfake detection, named Region-Based Optimization (RegO). Under the same acoustic environments, real audio exhibits a more compact feature distribution compared to fake audio (Ma et al. 2021; Zhang et al. 2023, 2024), so they can be seen as a whole from the same dataset. Based on this observation, we propose that for new tasks, gradient updates for real audio should be parallel to the gradient directions of previous tasks, while the updates for fake audio should be orthogonal to the previous task gra-

*Corresponding authors.

dients. Specifically, we utilize the Fisher information matrix (FIM) to measure the importance of neurons in the model, and then calculate the FIM for both real and fake audio detection separately. Following the fundamental principle of not constraining unimportant neurons to allow the model to quickly adapt to new tasks, and optimizing important neurons to overcome catastrophic forgetting, we divide the neurons into four regions for fine-grained region-adaptive gradient optimization: neurons that are unimportant for both real and fake audio detection are finetune directly to quickly adapt to new tasks; neurons that are important only for real audio detection are updated in parallel to the previous task gradients; neurons important only for fake audio detection are updated orthogonally to the previous task gradients; and neurons important for both real and fake audio detection are optimized through adaptive gradient updates based on the ratio of real to fake samples to maintain memory stability.

However, when the number of neurons in the model remains fixed as tasks increase, two problems arise: First, the number of neurons in less important regions decreases, making it progressively harder for the model to adapt to new tasks. Second, redundant neurons appear—those that are beneficial for only a few specific tasks but ineffective for others. To address these challenges, we draw inspiration from the non-linear nature of human memory forgetting (Loftus 1985; Ebbinghaus 2013). Over time, the memories that persist are generally those involving deeply understood knowledge, while other redundant Knowledge are forgotten. Based on this principle, we further propose a neuron forgetting mechanism inspired by the Ebbinghaus forgetting curve (Woźniak, Gorzelańczyk, and Murakowski 1995) to release redundant neurons from previous tasks. This mechanism enables the model to learn knowledge from new tasks more efficiently, ensuring quicker adaptation to new tasks and the acquisition of more generalized discriminative information.

Our experiments on the Evolving Deepfake Audio (EVDA) benchmark (Zhang, Yi, and Tao 2024) demonstrate that our method outperforms several mainstream continual learning methods and state-of-the-art continual audio deepfake detection methods, including RAWM and RWM, in terms of balancing stability and plasticity. Furthermore, our method can be easily extended to other domains. General study conducted on image recognition tasks also showed competitive results, highlighting the potential significance of our approach across different machine learning fields.

In summary, we make the following contributions:

- We propose a continual learning method for audio deepfake detection, named RegO, which partitions the neural network parameter space into four regions using the FIM. This method facilitates fine-grained, region-adaptive gradient optimization, ensuring an optimal trade-off between memory stability and learning plasticity.
- We further propose a neuron forgetting mechanism based on Ebbinghaus forgetting curve, which releases redundant neurons from previous tasks to ensure faster adaptation to new tasks and to learn more generalizable discriminative information.
- We conduct extensive experiments on the EVDA benchmark to validate the effectiveness of our method. Furthermore, we perform general study, and the results indicate that our approach holds potential significance in other domains, such as image recognition, without being limited to a specific field.

Related Work

Continual learning is a machine learning paradigm that aims to enable models to retain and use previously learned knowledge while continuously learning new tasks, thereby overcoming catastrophic forgetting. Current mainstream methods can be categorized into the following types: Regularization-based methods, which balance new and old tasks by selectively adding regularization terms to the changes in network parameters, such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al. 2017), Synaptic Intelligence (SI) (Zenke, Poole, and Ganguli 2017), Gradient Episodic Memory (GEM) (Lopez-Paz and Ranzato 2017), Orthogonal Weight Modification (OWM) (Zeng et al. 2019) etc. (Qiao et al. 2024; Elsayed and Mahmood 2023) Replay-based methods, which carefully select training samples from previous tasks, store them in a buffer, and mix them with new task training samples to ensure accuracy for old tasks, such as Greedy Sampler and Dumb Learner (GDumb) (Prabhu, Torr, and Dokania 2020) and CWRStar (Lomonaco et al. 2020); Architecture-based methods, which use parameter isolation and dynamic expansion of the parameter space to protect previously acquired knowledge. (Wang et al. 2022; Razdaibiedina et al. 2023)

Mainstream continual learning methods have achieved significant success in various fields, including image classification (Razdaibiedina et al. 2023; Yoo et al. 2024), object detection (Menezes et al. 2023), and semantic segmentation (Camuffo and Milani 2023; Zhu et al. 2023). Continual learning methods are also effective in audio deepfake detection, empowering models to recognize diverse and unseen fake audio by leveraging continual learning principles. Currently, most continual learning methods for audio deepfake detection are regularization-based, such as Detecting Fake Without Forgetting (DFWF) (Ma et al. 2021), RAWM (Zhang et al. 2023), and RWM (Zhang et al. 2024), and have demonstrated impressive results. However, the approximations made by regularization methods can lead to cumulative errors during continual learning, affecting the balance between model stability and plasticity. In contrast, our method selectively adjusts the gradients of parameters that are crucial for previous tasks while fine-tuning the remaining parameters to learn new tasks. This allows our method to adapt more quickly to new tasks while preventing catastrophic forgetting of old ones.

Methodology

In this section, the Region-Based Optimization (RegO) architecture, as illustrated in Figure 1, encompasses the principles and implementation of three core modules: Importance Region Localization (IRL), Region-Adaptive Optimization (RAO), and the Ebbinghaus Forgetting Mechanism (EFM).

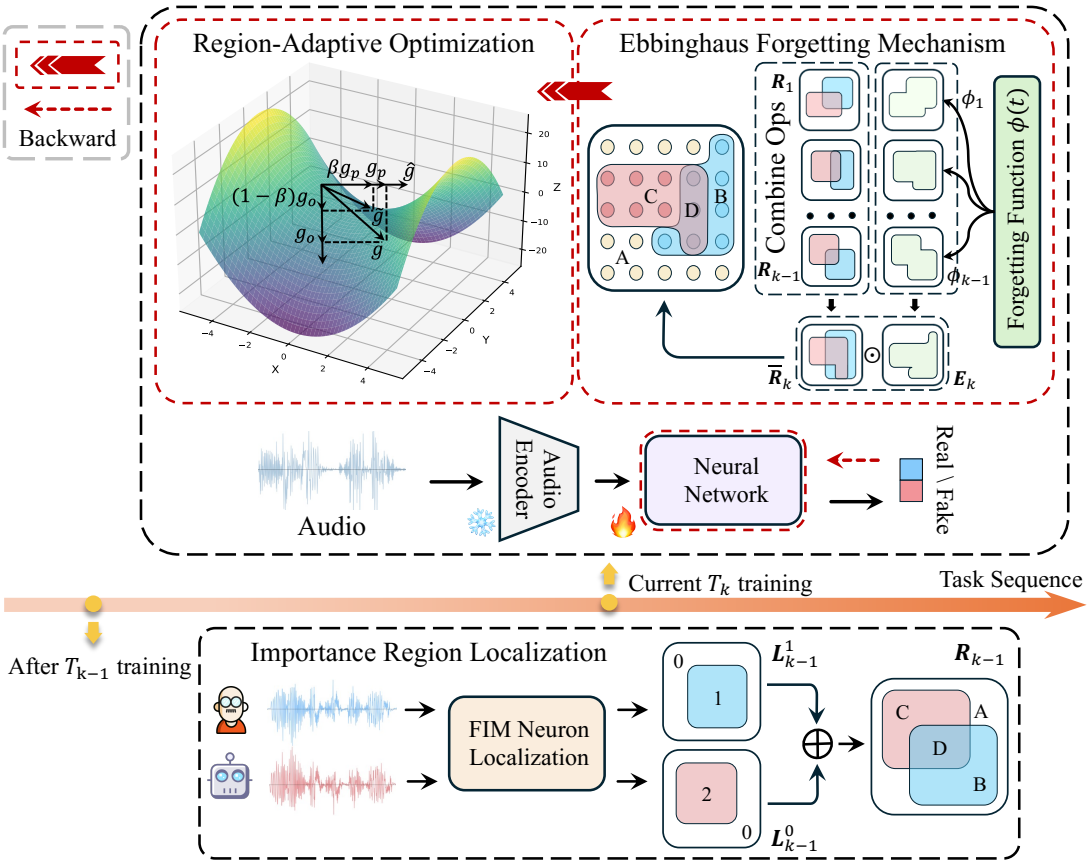

Figure 1: Illustration of RegO Architecture. (i).After the training of each task, we calculate the region matrix $\mathbf{R}$ through the IRL module. (ii).From the second task onward, gradient optimization is performed during backpropagation (shown by dark red arrows and boxes). The $\mathbf{E}_k$ is obtained through the EFM module and then combined with the historical $\mathbf{R}$ to generate $\overline{\mathbf{R}}$. (iii).Based on $\overline{\mathbf{R}}$, the RAO module updates the model weights as follows: Region A: fine-tuning (i.e. g); Region B: gradient update in the projection direction (i.e. g_p); Region C: gradient update in the orthogonal direction (i.e. g_o); Region D: adaptive gradient update based on the number of samples (i.e. $\widehat{g}$).

Definitions and Notation

In Continual Learning, we define a sequence of tasks $\{T_1, T_2, T_3, \ldots, T_N\}$ of N tasks. For the k-th task T_k, there is a corresponding training set $\mathcal{D}_k = \left\{\left(x_k^i, y_k^i\right)\right\}_{i=1}^{N_k}$ and corresponding parameters θ_k. The prediction of the model on input x is denoted by $f(x;\theta_k)$. During the training of task k, we define the loss function as follows:

$$\mathcal{L}(\theta_k, \mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|} \sum_{(x,y)\in\mathcal{D}_k} CE(f(x;\theta_k), y) \quad (1)$$

where CE is the standard cross-entropy loss. The gradient is shown as follows:

$$g_k = \nabla_{\theta_k} \mathcal{L}(\theta_k, D_k) \quad (2)$$

Importance Region Localization

First, we need to identify the neuron regions that are important for the previous tasks. Inspired by (Kirkpatrick et al. 2017; Huszár 2018), we choose the Fisher Information Matrix (FIM) as a measure of neuron importance. After completing the training of the k-th task, we pass real and fake audio through the model separately to calculate the corresponding FIM, which encapsulate the importance measures of the weights for both real and fake audio detection. The FIM is defined as follows:

$$\mathbf{F}_k^{(c)} = \mathbb{E}\left[\nabla_\theta \log p(D_k^{(c)}|\theta) \nabla_\theta \log p(D_k^{(c)}|\theta)^\top \Big|_{\theta=\theta_k^*}\right] \quad (3)$$

Here, c is 0 (fake) or 1 (real), D_k represents the training dataset corresponding to T_k, and θ_k^* denotes the optimal parameters for T_k. Note that the log probability of the data $D_k^{(c)}$ given the parameters $\log p(D_k^{(c)}|\theta)$ is simply the negative of the loss function $-\mathcal{L}(\theta_k, \mathcal{D}_k)$ for task T_k. Based on this, by setting the threshold α, we locate important neurons, resulting in a localization matrix, as shown in Equation 4.

$$\mathbf{L}_k^{(c)}[i][j] = \begin{cases} 2 & \text{if } c=0 \text{ and } \mathbf{F}_k^{(c)}[i][j] \geq \mathrm{P}_\alpha(\mathbf{F}_k^{(c)}) \\ 1 & \text{if } c=1 \text{ and } \mathbf{F}_k^{(c)}[i][j] \geq \mathrm{P}_\alpha(\mathbf{F}_k^{(c)}) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where i, j represents the neuron index, P_α represents the α-percentile. Based on this, by summing the localization matri-

ces for both real and fake audio, we can identify four distinct regions, as shown in Equation 5.

$$\mathbf{R}_k[i][j] = \mathbf{L}_k^0[i][j] + \mathbf{L}_k^1[i][j] \tag{5}$$

In the $\mathbf{R}_k$, regions with values 0, 1, 2, and 3 are denoted by the letters A, B, C, and D, respectively. Region A represents neurons that are not important for T_k. Region B represents neurons that are important for real audio detection in T_k. Region C represents neurons that are important for fake audio detection in T_k. Region D represents neurons that are important for both real and fake audio detection in T_k.

Region-Adaptive Optimization

During training, starting from the second task, we merge all region matrices $\mathbf{R}$ from the previous tasks to obtain $\overline{\mathbf{R}}$. For the four regions A, B, C, and D within $\overline{\mathbf{R}}$, we adaptively optimize each region according to the following principles.

Firstly, because the neurons in region A have minimal impact on previous tasks, but are likely to become important for new tasks, we allow them to quickly adapt and learn the knowledge of the new tasks. Therefore, we do not apply additional gradient optimization to the neurons in region A and instead update them using fine-tuning, the gradient of region A is defined in Equation 6.

$$g_A = g \odot \mathbb{I}_{\{\overline{\mathbf{R}}[i][j]=0\}} \tag{6}$$

where $\odot$ denote the Hadamard product, $\mathbb{I}_{\{\overline{\mathbf{R}}[i][j]=0\}}$ is an indicator function that takes a value of 1 when $\overline{\mathbf{R}}[i][j]$ equals 0 and 0 otherwise.

Second, as mentioned above, real audio has a more compact feature distribution compared to fake audio and can be considered as coming from the same dataset. Therefore, to largely retain the knowledge of real audio detection from previous tasks, the neurons in region B should project the current gradient g onto the direction of the old task gradient $\hat{g}$ for gradient optimization, as shown in Equation 7.

$$\begin{aligned} g_p &= \frac{g \cdot \hat{g}}{\|\hat{g}\|^2}\hat{g} \\ g_B &= g_p \odot \mathbb{I}_{\{\overline{\mathbf{R}}[i][j]=1\}} \end{aligned} \tag{7}$$

Thirdly, due to the diversity of speech synthesis and voice conversion methods, there is a wide variance in feature distributions among fake audio samples. Therefore, to reduce the forgetting of discriminative information about fake audio from previous tasks while learning discriminative information for fake audio in new tasks, we update the gradient direction of neurons in region C to be orthogonal to the gradient direction of the old tasks, as shown in Equation 8.

$$\begin{aligned} g_o &= g - g_p \\ g_C &= g_o \odot \mathbb{I}_{\{\overline{\mathbf{R}}[i][j]=2\}} \end{aligned} \tag{8}$$

Fourth, region D is crucial for both real and fake audio discrimination from previous tasks. Therefore, to balance the retention of knowledge for both real and fake audio detection in neurons of region D, we need to optimize the gradient update direction to achieve an optimal trade-off.

Algorithm 1: Region-Based Optimization

Require: Training data from different datasets, η (learning rate), $\mathcal{R}$ (Region matrix set)

for every dataset k **do**
  for every batch b **do**
    if k = 1 **then**
      Update θ_k: $\theta_k \leftarrow \theta_k - \eta w$
    else
      Compute memory matrix $\mathbf{M}_k$ by Equ.(13)
      Compute Ebbinghaus matrix $\mathbf{E}_k$ by Equ.(14)
      Compute $\overline{\mathbf{R}}_k$ by combining region matrix set $\mathcal{R}$
      $g_A \leftarrow g \odot \mathbb{I}_{\{\overline{\mathbf{R}}_k[i][j]=0\}}$
      $g_p \leftarrow \frac{g \cdot \hat{g}}{\|\hat{g}\|^2}\hat{g}$
      $g_B \leftarrow g_p \odot \mathbb{I}_{\{\overline{\mathbf{R}}_k[i][j]=1\}}$
      $g_o \leftarrow g - g_p$
      $g_C \leftarrow g_o \odot \mathbb{I}_{\{\overline{\mathbf{R}}_k[i][j]=2\}}$
      $\beta \leftarrow \frac{\sum_{l=1}^{u} N^l}{\sum_{l=1}^{u+v} N^l}$
      $\widetilde{g} \leftarrow \beta * g_p + (1-\beta) * g_o$
      $g_D \leftarrow \widetilde{g} \odot \mathbb{I}_{\{\overline{\mathbf{R}}_k[i][j]=3\}}$
      Initialization: $w \leftarrow 0$
      $w \leftarrow g_A + g_B + g_C + g_D$
      Update θ_k: $\theta_k \leftarrow \theta_k - \eta w$
    end if
  end for
  Compute the k-th Region Matrix $\mathbf{R}_k$ by Equ.(1)(2)(3)
  $\mathcal{R} \leftarrow \mathbf{R}_k$
end for

Specifically, we adaptively determine whether the gradient update direction should lean more towards the projection direction or the orthogonal direction based on the proportion of real and fake audio samples, as shown in Equations 9 and 10.

$$\beta = \frac{\sum_{l=1}^{u} N^l}{\sum_{l=1}^{u+v} N^l} \tag{9}$$

$$\begin{aligned} \widetilde{g} &= \beta * g_p + (1-\beta) * g_o \\ g_D &= \widetilde{g} \odot \mathbb{I}_{\{\overline{\mathbf{R}}[i][j]=3\}} \end{aligned} \tag{10}$$

In Equation 9, u and v represent the number of classes with similar feature distributions and the remaining classes, respectively. In deepfake audio detection, u and v are both set to 1, indicating the two classes of real and fake audio. In image recognition, u and v represent the number of classes with similar feature distributions and the number of classes with dissimilar feature distributions, respectively. N^l denotes the number of samples in the batch for the l-th class.

Finally, the total gradient update for a batch is defined as:

$$w = g_A + g_B + g_C + g_D \tag{11}$$

Ebbinghaus Forgetting Mechanism

During the continual learning process, when the number of neurons remains constant, neurons in region A gradually diminish, and redundant neurons that only benefit individual tasks begin to emerge, which undermines the model's

Continual Learning Methods	EER (↓) on each experience								
	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7	Exp_8	Avg
Replay-All	2.80	5.68	1.52	0.76	1.84	7.96	5.76	2.56	3.61
Finetune-Exp_1	2.20	24.80	23.16	16.84	23.80	34.12	26.44	15.52	20.86
Finetune	5.16	15.56	8.20	2.32	4.08	21.72	9.64	3.04	8.72
EWC	**3.72**	13.92	7.32	2.12	3.56	17.40	10.24	3.16	7.68
GDumb	4.72	14.12	7.32	4.60	6.56	24.28	15.28	11.40	11.03
GEM	5.60	16.56	6.28	2.60	9.60	24.44	11.88	4.28	10.15
CWRStar	5.12	27.92	22.88	29.36	45.52	43.20	49.92	18.32	30.28
SI	6.96	10.88	5.92	1.60	4.04	18.96	10.04	3.32	7.71
OWM	27.28	33.72	29.32	33.12	47.28	49.52	48.80	26.32	36.92
RAWM	9.28	16.04	6.76	2.60	3.60	19.52	9.64	3.40	8.85
RWM	4.44	14.92	6.28	1.92	4.44	18.92	10.04	3.52	8.06
RegO (Ours)	4.36	**10.64**	**3.76**	**1.20**	**3.16**	**15.72**	**9.16**	**2.72**	**6.34**

Table 1: The EER (%) of our method compared with various methods.

adaptability and generalization ability. To address this issue, inspired by Ebbinghaus forgetting theory (Loftus 1985; Ebbinghaus 2013), we propose a neuron forgetting mechanism based on the Ebbinghaus memory curve (Woźniak, Gorzelańczyk, and Murakowski 1995). The approximation function is defined as Equation 12.

$$\phi(t) = e^{-\frac{t}{k}} \tag{12}$$

where t represents the time steps and k denote the k-th task. Specifically, we define the Ebbinghaus forgetting curve function based on the number of tasks processed so far, and calculate the memory weights using this forgetting curve function. Then, we allocate memory weights to the region matrix $\mathbf{R}$ of the old tasks. By Equation 13 and 14, we compute the memory matrix $\mathbf{M}$, which contains the accumulated memory weights for each neuron. Finally, we set a threshold γ. When the memory weight of a neuron is less than γ, that neuron is released, resulting in the Ebbinghaus matrix $\mathbf{E}$.

$$\mathbf{M}_k[i][j] = \sum_{t=1}^{k-1} \phi(t) * \mathbb{I}_{\{\mathbf{R}_t[i][j] \in \{0,1,2,3\}\}} \tag{13}$$

$$\mathbf{E}_k[i][j] = \begin{cases} 1 & \text{if } \mathbf{M}_k[i][j] > \gamma \\ 0 & \text{if } \mathbf{M}_k[i][j] \leq \gamma \end{cases} \tag{14}$$

Experiments

We conduct a series of experiments to evaluate the effectiveness of our approach. The experiments are performed on a continual learning benchmark EVDA (Zhang, Yi, and Tao 2024) for speech deepfake detection, which includes eight publicly available and popular datasets specifically designed for incremental synthesis algorithm audio deepfake detection. Additionally, we carry out a general study in the field of image recognition using the well-established continual learning benchmark, CLEAR (Lin et al. 2021).

Experimental Setup

Datasets and Metrics In this paper, we refer to each dataset as "Exp" (e.g., Exp_1, Exp_2, ..., Exp_{10}), representing the different datasets used in our experiments. The EVDA benchmark from Exp_1 to Exp_8 are FMFCC (Zhang et al. 2021), In the Wild (Müller et al. 2022), ADD 2022 (Yi et al. 2022), ASVspoof2015 (Wu et al. 2017), ASVspoof2019 (Todisco et al. 2019), ASVspoof2021 (Yamagishi et al. 2021), FoR (Reimao and Tzerpos 2019), and HAD (Yi et al. 2021). For the EVDA baseline, 2000 samples are randomly sampled from each dataset as the training set, and 5000 samples are sampled as the test set. The EVDA baseline dataset configuration includes cross-lingual (Chinese and English) and cross-task (whole-segment and partial-segment fake detection) scenarios to simulate the unseen and diverse real-world forgery conditions. The final model in this study refers to the model trained sequentially on these eight datasets and evaluated on each dataset. We use the standard metric Equal Error Rate (EER) (Wu et al. 2017) in the field of audio deepfake detection to evaluate the performance of our model.

Model We employ the pre-trained speech model Wav2vec 2.0 (Baevski et al. 2020) as the feature extractor, the parameters of Wav2vec 2.0 are loaded from the pretrained model XLSR-53 (Conneau et al. 2021). Given the robustness of the speech features obtained from the pre-trained model, we opt for a 5-layer SimpleMlp as the backend, which consists of fully connected layers with the following dimensions: 1024 to 512, 512 to 512 (x3), and 512 to 2. The code is available at https://github.com/cyjie429/RegO.

Training Details We use the Adam optimizer to finetune the SimpleMlp, with a learning rate η of 0.0001 and a batch size of 32, on an NVIDIA A100 80GB GPU. To evaluate the performance of our method, we compare it with six widely used continual learning methods, finetuning, and two advanced continual learning methods specifically designed for audio deepfake detection: RAWM (Zhang et al. 2023), RWM (Zhang et al. 2024). Additionally, we present the training results on all datasets (Replay-All), which are considered the lower bound for all mentioned continual learning methods (Parisi et al. 2019).

Ablation Study	EER (↓) on each experience								
	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7	Exp_8	Avg
RegO (Ours)	**4.36**	10.64	**3.76**	**1.20**	3.16	**15.72**	**9.16**	**2.72**	**6.34**
w/o EFM	4.48	**10.48**	5.56	1.32	**3.12**	16.48	9.68	3.32	6.80
w/o IRL	5.32	12.28	4.72	1.52	3.48	19.08	9.76	2.92	7.38
w/o RAO	4.68	14.68	5.52	2.48	4.44	16.16	9.80	3.32	7.63

Table 2: The EER (%) results of the ablation study for our method.

Comparison with Other Methods

In this experiment, we compare RegO with other methods. Here, Finetune-Exp_1 shows the results of training on Exp_1 and evaluating on the other Exps, highlighting the significant differences between the various Exps. As shown in Table 1, after training on 8 Exps, our method achieves the best performance on 7 of the Exps and the second-best performance on Exp1. The overall evaluation metrics demonstrate that our method is only slightly inferior compared to Replay-ALL, which is considered the upper bound for continual learning performance. Additionally, among the eight Exps, Exp_1, Exp_3, and Exp_8 are Chinese datasets, while the remaining ones are English datasets. Notably, Exp_3 consists of low-quality speech data, and Exp_8 is a partial-fake spoofing dataset. The experimental results demonstrate that our method shows promising potential in both cross-lingual and cross-task scenarios, indicating its capability to handle relatively diverse real-world audio deepfake environments.

Ablation Study

We conduct an ablation study to evaluate the effectiveness of the proposed modules. The results, shown in Table 3, are as follows: "w/o EFM" denotes the removal of the Ebbinghaus forgetting mechanism, "w/o IRL" indicates no division between real and fake regions, and "w/o RAO" refers to applying orthogonal gradient optimization to all weights. The results for "w/o RAO" suggest that, compared to orthogonally optimizing all weights, applying region-adaptive gradient optimization to critical regions alone achieves a better balance between model memory stability and learning plasticity. Additionally, we observe that "w/o EFM" performs better than RegO on some of the earlier datasets, but its overall capability is inferior to RegO. We attribute this to the significant differences among the eight audio deepfake detection datasets, which result in a substantial number of redundant neurons. The EFM module effectively filters out these redundant neurons, enabling the model to adapt more quickly to other tasks.

Hyperparameter Study

We conduct a hyperparameter study to evaluate the impact of α and γ on our method RegO. Notably, the α study is performed with γ fixed at 0.1, while the γ study is conducted with α fixed at 0.75. The experimental results show that our method performs best when α is set to 0.75. As α increases, the region of important neurons (i.e., regions B, C, and D) shrinks, reducing model stability. Conversely, as α decreases, the region of less important neurons (i.e., region A) diminishes, leading to reduced model plasticity. Both scenarios result in a decline in model performance. In the γ experiments, increasing γ causes more neurons to be classified as redundant, including important neurons effective across multiple early tasks that are mistakenly classified as redundant. This misclassification reduces model stability, leading to a decline in performance.

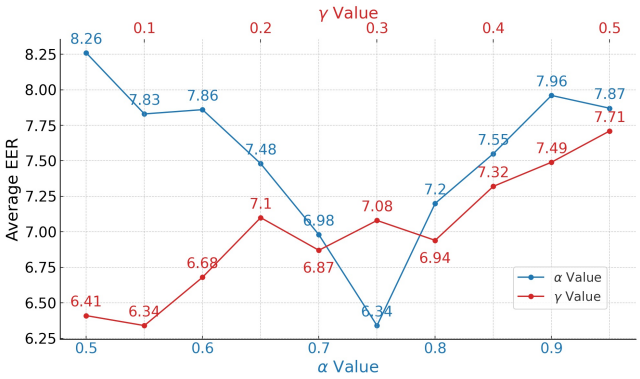

Figure 2: The average EER (%) results of the hyperparameter study for our method RegO.

General Study

Experimental Setup We use the CLEAR benchmark (Lin et al. 2021) to evaluate the scalability of our method. CLEAR is a classic continual image classification benchmark, with datasets based on the natural temporal evolution of visual concepts in the real world. It adopts task-based sequential learning by dividing the temporal stream into 10 buckets, each composed of labeled subsets for training and evaluation (with 300 images in the training set and 150 in the test set), resulting in a series of 11-way classification tasks. A small labeled subset (Exp_1, Exp_2, ..., Exp_{10}) consists of 11 temporally dynamic categories, including examples like computers, cosplay, etc. We use classification accuracy to evaluate model performance. For the image recognition model, we use a pre-trained ResNet-50 (He et al. 2016) as the feature extractor, which is frozen during continual learning, generating 2048-dimensional features. The downstream classifier has three linear layers: 2048 to 1024, 1024 to 512, and 512 to 2. We set the initial learning rate to 0.1, a batch size of 512, and used SGD optimizer with 0.9 momentum.

Continual Learning Methods	Accuracy (↑) on each experience									
	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7	Exp_8	Exp_9	Exp_{10}
Replay-All	94.34	94.44	94.44	94.85	95.66	94.14	93.94	95.86	94.24	95.56
Finetune-Exp_1	57.27	56.67	59.60	58.89	59.39	55.05	56.16	54.75	54.65	55.15
Finetune	90.40	89.80	90.10	92.73	90.71	90.40	90.10	89.90	90.40	92.42
EWC	91.72	91.62	91.31	92.12	91.31	90.40	91.11	90.61	90.71	93.13
GEM	91.62	90.51	90.30	92.93	91.62	89.39	90.30	90.10	89.49	93.33
GDumb	90.20	87.78	89.60	89.09	89.19	86.57	87.47	88.18	87.88	88.38
CWRStar	87.68	87.98	87.58	88.79	88.79	86.77	87.58	86.77	86.77	90.00
SI	89.39	89.29	90.00	91.11	89.79	88.69	89.39	89.90	88.79	90.40
OWM	73.03	71.41	70.30	73.13	71.01	68.99	69.70	67.27	70.30	69.70
RAWM	85.25	84.95	82.83	83.84	84.14	81.62	81.52	83.64	83.84	82.42
RWM	87.17	86.26	87.68	89.29	87.17	85.66	88.18	85.15	86.87	85.86
RegO (Ours)	**91.92**	**93.03**	**92.63**	**93.64**	**93.94**	**92.32**	**92.53**	**92.53**	**92.42**	**94.75**

Table 3: The accuracy (%) of the models trained on all CLEAR experiments. All results are reproduced by us.

Ablation Study	Accuracy (↑) on each experience									
	Exp_1	Exp_2	Exp_3	Exp_4	Exp_5	Exp_6	Exp_7	Exp_8	Exp_9	Exp_{10}
RegO (Ours)	91.92	93.03	92.63	93.64	93.94	**92.32**	92.53	92.53	92.42	94.75
w/o EFM	**92.93**	**93.93**	**93.33**	**94.65**	**94.34**	91.82	**93.23**	**93.64**	**93.13**	**94.85**
w/o IRL	90.61	89.70	90.61	91.21	91.41	88.89	90.71	89.49	89.49	92.83
w/o RAO	88.59	87.37	88.48	89.49	88.99	87.98	87.67	88.48	87.37	90.91

Table 4: The accuracy (%) results of the ablation study for our method on the CLEAR experiences.

Comparison with Other Methods We compare RegO with several classic continual learning methods. As shown in Table 3, after training on 10 Exps, the performance of RegO is second only to Replay-All, which is considered the upper bound for continual learning performance. Additionally, Table 3 shows that RWM and RAWM perform better in the earlier subset (Exp_1 - Exp_5) compared to the later ones, indicating that these algorithms are more focused on mitigating forgetting, but are less adaptable to new tasks. Our method overcomes catastrophic forgetting by optimizing the gradients of important neurons, while fine-tuning less critical neurons directly to ensure rapid adaptation to new tasks, ensuring an appropriate stability-plasticity trade-off.

Ablation Study We conduct an ablation study to evaluate the effectiveness of the proposed modules, with the results shown in Table 4. Compared to Table 3, we observe an interesting phenomenon: in the image recognition task, removing the EFM module leads to better model performance, which contrasts with the ablation results for audio deepfake detection. We speculate that this is because the CLEAR benchmark represents a decade-long natural temporal evolution of real-world visual concepts, where the appearance of major categories such as computers, cameras, etc. has remained relatively unchanged over the years. The results of Finetune-Exp_1 support this hypothesis: after training on Exp_1, the accuracy remains similar from Exp_1 to Exp_5 but shows a slight decline from Exp_5 to Exp_{10}. This indicates that retaining old knowledge might interfere with performance across multiple tasks. On the other hand, in the audio deepfake detection task, where differences in synthesis or conversion algorithms are more distinct (as shown by the Finetune-Exp_1 results in Table 1), the role of EFM becomes more critical. Nevertheless, regardless of whether the EFM is integrated, both versions of our method consistently outperform other methods.

Conclusion

In this paper, we propose an effective continual learning algorithm, Region-Based Optimization (RegO), aimed at improving the generalization of audio deepfake detection models against diverse and unseen forgeries in real-world scenarios. The core idea of RegO is to avoid constraints on less important neurons, allowing the model to quickly adapt to new tasks, while applying region-based adaptive gradient optimization to important neurons to overcome catastrophic forgetting, achieving a suitable balance between memory stability and learning plasticity. Experimental results demonstrate that our method outperforms SOTA method RWM for audio deepfake detection, proving its robustness against diverse forgery techniques. Additionally, we conduct the general study and achieve competitive results, indicating our method has potential significance in other domains, such as image recognition. Moreover, we plan to explore how to extend our method to address other challenges in machine learning, such as multi-task learning (Langa 2021).

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) (No.62322120, No.U21B2010, No.62306316, No.62206278, No.62201002, 6247077204), the STI 2030—Major Projects (No. 2021ZD0201500), Excellent Youth Foundation of Anhui Scientific Committee (No. 2408085Y034), Distinguished Youth Foundation of Anhui Scientific Committee (No. 2208085J05), Special Fund for Key Program of Science and Technology of Anhui Province (No. 202203a07020008), Cloud Ginger XR-1.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Camuffo, E.; and Milani, S. 2023. Continual Learning for LiDAR Semantic Segmentation: Class-Incremental and Coarse-to-Fine strategies on Sparse Data. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2447–2456.

Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; and Auli, M. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, 2426–2430.

Ebbinghaus, H. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4): 155.

Elsayed, M.; and Mahmood, A. R. 2023. Addressing Loss of Plasticity and Catastrophic Forgetting in Continual Learning. In *The Twelfth International Conference on Learning Representations*.

Fan, C.; Dong, S.; Xue, J.; Chen, Y.; Yi, J.; and Lv, Z. 2024. Frequency-mix Knowledge Distillation for Fake Speech Detection. *arXiv preprint arXiv:2406.09664*.

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

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.

Huszár, F. 2018. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences*, 115(11).

Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; and Lee, K. A. 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Interspeech 2017*, 2–6. International Speech Communication Association.

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

Langa, J. 2021. Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101: 761.

Liavas, A.; and Regalia, P. 1999. On the numerical stability and accuracy of the conventional recursive least squares algorithm. *IEEE Transactions on Signal Processing*, 47(1): 88–96.

Lin, Z.; Shi, J.; Pathak, D.; and Ramanan, D. 2021. The clear benchmark: Continual learning on real-world imagery. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

Loftus, G. R. 1985. Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(2): 397.

Lomonaco, V.; Maltoni, D.; Pellegrini, L.; et al. 2020. Rehearsal-Free Continual Learning over Small Non-IID Batches. In *CVPR Workshops*, 2, 3.

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

Ma, H.; Yi, J.; Tao, J.; Bai, Y.; Tian, Z.; and Wang, C. 2021. Continual Learning for Fake Audio Detection. In *Proc. Interspeech 2021*, 886–890.

Menezes, A. G.; de Moura, G.; Alves, C.; and de Carvalho, A. C. 2023. Continual Object Detection: A review of definitions, strategies, and challenges. *Neural Networks*, 161: 476–493.

Müller, N.; Czempin, P.; Diekmann, F.; Froghyar, A.; and Böttinger, K. 2022. Does Audio Deepfake Detection Generalize? In *Proc. Interspeech 2022*, 2783–2787.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113: 54–71.

Prabhu, A.; Torr, P. H.; and Dokania, P. K. 2020. Gdumb: A simple approach that questions our progress in continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 524–540. Springer.

Qiao, J.; Tan, X.; Chen, C.; Qu, Y.; Peng, Y.; Xie, Y.; et al. 2024. Prompt Gradient Projection for Continual Learning. In *The Twelfth International Conference on Learning Representations*.

Razdaibiedina, A.; Mao, Y.; Hou, R.; Khabsa, M.; Lewis, M.; and Almahairi, A. 2023. Progressive Prompts: Continual Learning for Language Models. In *The Eleventh International Conference on Learning Representations*.

Reimao, R.; and Tzerpos, V. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 1–10. IEEE.

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

Tak, H.; Todisco, M.; Wang, X.; Jung, J.-w.; Yamagishi, J.; and Evans, N. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The Speaker and Language Recognition Workshop*.

Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; and Lee, K. A. 2019. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *Interspeech 2019*, 1008–1012. International Speech Communication Association.

Wang, C.; Yi, J.; Tao, J.; Zhang, C. Y.; Zhang, S.; and Chen, X. 2023. Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features. In *Proc. INTERSPEECH 2023*, 3844–3848.

Wang, L.; Zhang, X.; Li, Q.; Zhu, J.; and Zhong, Y. 2022. Coscl: Cooperation of small continual learners is stronger than a big one. In *European Conference on Computer Vision*, 254–271. Springer.

Wang, X.; and Yamagishi, J. 2022. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 100–106.

Woźniak, P.; Gorzelańczyk, E.; and Murakowski, J. 1995. Two components of long-term memory. *Acta neurobiologiae experimentalis*, 55(4): 301–305.

Wu, Z.; Yamagishi, J.; Kinnunen, T.; Hanilçi, C.; Sahidullah, M.; Sizov, A.; Evans, N.; Todisco, M.; and Delgado, H. 2017. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing*, 11(4): 588–604.

Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K. A.; Kinnunen, T.; Evans, N.; et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*.

Yi, J.; Bai, Y.; Tao, J.; Ma, H.; Tian, Z.; Wang, C.; Wang, T.; and Fu, R. 2021. Half-Truth: A Partially Fake Audio Detection Dataset. In *Proc. Interspeech 2021*, 1654–1658.

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

Yi, J.; Tao, J.; Fu, R.; Yan, X.; Wang, C.; Wang, T.; Zhang, C. Y.; Zhang, X.; Zhao, Y.; Ren, Y.; et al. 2023a. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C. Y.; and Zhao, Y. 2023b. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

Yoo, J.; Liu, Y.; Wood, F.; and Pleiss, G. 2024. Layerwise Proximal Replay: A Proximal Point Method for Online Continual Learning. In *Forty-first International Conference on Machine Learning*.

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

Zenke, F.; Poole, B.; and Ganguli, S. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, 3987–3995. PMLR.

Zhang, X.; Yi, J.; and Tao, J. 2024. EVDA: Evolving Deepfake Audio Detection Continual Learning Benchmark. *arXiv preprint arXiv:2405.08596*.

Zhang, X.; Yi, J.; Tao, J.; Wang, C.; and Zhang, C. Y. 2023. Do you remember? overcoming catastrophic forgetting for fake audio detection. In *International Conference on Machine Learning*, 41819–41831. PMLR.

Zhang, X.; Yi, J.; Wang, C.; Zhang, C. Y.; Zeng, S.; and Tao, J. 2024. What to remember: Self-adaptive continual learning for audio deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19569–19577.

Zhang, Z.; Gu, Y.; Yi, X.; and Zhao, X. 2021. FMFCC-a: a challenging Mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*, 117–131. Springer.

Zhu, L.; Chen, T.; Yin, J.; See, S.; and Liu, J. 2023. Continual Semantic Segmentation with Automatic Memory Sample Selection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3082–3092.