ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446846

IMPROVING CONTINUAL LEARNING OF ACOUSTIC SCENE CLASSIFICATION VIA MUTUAL INFORMATION OPTIMIZATION

Muqiao Yang[1], *Umberto Cappellazzo*[2], *Xiang Li*[1], *Bhiksha Raj*[1]

[1]Carnegie Mellon University, [2]University of Trento

ABSTRACT

Continual learning, which aims to incrementally accumulate knowledge, has been an increasingly significant but challenging research topic for deep models that are prone to catastrophic forgetting. In this paper, we propose a novel replay-based continual learning approach in the context of class-incremental learning in acoustic scene classification, to classify audio recordings into an expanding set of classes that characterize the acoustic scenes. Our approach is improving both the modeling and memory selection mechanism via mutual information optimization in continual learning. By regarding incremental classes of acoustic scenes as different tasks, our model is expected to learn both task-agnostic and task-specific knowledge by replaying representative and informative samples. This optimization also enables the model to utilize past knowledge effectively and learn from new information during continual learning. We demonstrate that our approach has a superior performance compared to existing methods on multiple datasets and continual learning evaluation metrics.

Index Terms— Continual learning, acoustic scene classification

1. INTRODUCTION

Acoustic scene classification (ASC) refers to the ability of humans or machines to recognize an environmental sound in a set of acoustic scene classes [1]. Even though the performance of deep neural network-based models has outperformed humans on this specific task [2], the success of existing ASC models has mostly relied on training with a large *fixed* set of data and *pre-defined* acoustic scene classes [1,3,4]. In contrast, humans are intrinsically capable of learning from streams of data with unseen classes, so that they can adapt themselves to the complex environment and evolve their knowledge in a lifelong manner [5]. Such an ability is referred to as continual learning, which is nontrivial for artificial intelligent systems and machine learning models.

The objective of continual learning is to address catastrophic forgetting [6], which means the tendency of models to abruptly erase past knowledge while learning new tasks. By a) introducing regularization terms to the loss functions [6, 7], b) separately learning task-specific knowledge with different modules of parameters [8, 9], or c) recovering the past data distribution by storing and replaying the memory [10–13], continual learning approaches have shown a strong performance in the scenario of dynamic data distributions. Under this practical setting, non-stationary data and tasks arrive in sequential order, but the model has limited or no access to the previous data that it has learned from [14].

From the perspective of human perception, recognizing an unseen acoustic scene can be regarded as an ensemble of background noises and sound events [15]. Then humans can associate new sound categories with specific acoustic scenes based on their extensive life experiences. Furthermore, Hendrickso et al. [16] showed that the perception of acoustic scenes in human brains is organized in terms of a general feature match between sounds and the referents in their memory. Therefore, we base our ASC model on the replay-based continual learning approaches, which save a small subset of past data into a memory buffer and replay samples of the memory when it encounters new tasks in the subsequent training process.

Specifically, in this work, we focus on the underexplored domain of continual learning in acoustic scene classification, and propose to improve both the training process and the memory selection procedure in our ASC model from the perspective of mutual information (MI) optimization. MI quantifies the amount of information between two variables, and has been proven to be helpful in representation learning from an information-theoretic view [17–21]. We base the optimization of MI on an architecture with a feature extractor followed by a classifier. Since feature extraction is to disentangle many distinct but informative factors from the data [22], we would like the feature extractor to learn generalized representations of the input audio that is agnostic to acoustic scene classes. Meanwhile, we would like the classifier to learn to map the extracted representations to correct classes. This paper demonstrates that mutual information can help the feature extractor learn *task-agnostic* knowledge, while helping the classifier learn *task-specific* knowledge. For the feature extractor part, we first present that it is theoretically sound to learn task-agnostic knowledge by maximizing the MI between the feature representations of the original input and an augmented acoustic scene of the same input. For the classifier part, we show that by selecting the memory samples with a combination of surprise and learnability criteria, the samples are expected to be both representative and informative to boost the continual learning performance of the ASC model. We evaluate our method on multiple datasets and continual learning metrics, showing that it can not only decrease the forgetting effect by learning task-specific knowledge, but improve the generalizability of the model by learning task-agnostic knowledge as well.

2. RELATED WORK

Replay-based, or rehearsal-based approaches, are a family of continual learning methods by providing old samples to approximate the past data distribution, with the price of a small-size memory buffer [10–13, 23–25]. To retrieve effective samples from the buffer to benefit the continual learning paradigm, various selection strategies have been investigated. Given a memory buffer with a fixed size, other than naively saving and replacing samples in the memory with an equal probability, reservoir sampling has been applied to compute how likely a sample is to be saved before the total size of a single pass of data is known [23]. Rebuffi et al. [24] prioritized the memory selection based on a herding mechanism, which favors selecting the samples whose features are the closest to the center of their class. Gradient-based sample selection (GSS) methods aim to increase the diversity of the samples in the memory by computing the cosine similarity of each sample with other random samples in the buffer through their gradients [25]. GSS methods are generally computationally costly because of the gradient computation.

Although most existing research on continual learning is concentrated on the domain of computer vision and reinforcement learning [6, 10, 11], there has also been a growing body of work starting to explore its application to speech-related tasks. Meanwhile, most of them are based on replaying approaches. Yang et al. [26] applied a gradient-based method to store an episodic memory in automatic speech recognition (ASR) where there are no explicit task boundaries. Xiao et al. [27] proposed to update the memory by selecting the samples based on the uncertainty of the sample embeddings through the inference of the ASC model. Cappellazzo et al. [28] introduced knowledge distillation in both the feature embedding space and the prediction space of spoken language understanding (SLU) models with a rehearsal memory. Our work differs from the replay-based methods described above by optimizing the modeling and memory selection for ASC in terms of MI.

There have also been some other algorithms trying to improve the continual learning performances from an information-theoretic perspective. Sun et al. [29] proposed to combine both surprise and learnability as an online memory selection criterion, such that the selected samples can be informative and avoid outliers. Guo et al. [30] provided a mutual information maximization method to encourage the online continual learning system to learn a holistic representation. Shi et al. [31] tried to preserve the information gain on model parameters by updating the parameters at a bit level, such that the loss of information gain can be reduced. Unlike prior work, our method shows that the optimization of mutual information can *a*) learn task-agnostic knowledge through augmented acoustic scenes and *b*) learn task-specific knowledge by selecting representative and informative samples from the memory buffer.

3. METHOD

3.1. Problem Statement

We follow the setting of class-incremental learning (CIL), where new classes of acoustic scenes keep appearing in continuous streams of data. Compared to another category of continual learning, i.e., task-incremental learning (TIL), CIL does not have access to task identities during inference time. Therefore, its objective is to build a holistic classifier among all of the seen classes by making use of the label information only.

In our context, a task is defined by a set of train and test data that follows a similar distribution, and in practice, it usually refers to a new set of data that contains data in different classes. For example, one task may consist of acoustic scenes from various vehicles, and another task may include sound events from animals. Consider we have the data streams $\mathcal{X} = \{X_t\}_{t=1}^{T}$ and its corresponding labels $\{Y_t^{\text{gt}}\}_{t=1}^{T}$, where X_t indicates the data at task t and T is the total number of tasks. Note that $Y_{t_i}^{\text{gt}} \cap Y_{t_j}^{\text{gt}} = \varnothing$ for all $t_i \neq t_j$ under this setting. For the modeling, we use an architecture of a feature extractor Φ and a classifier Θ, where the output of the feature extractor is $\Phi(X) = Z$, which is the feature representation of X and also the input of the classifier such that $\Theta(Z) = Y$, where Y indicates the predicted logits.

At the end of each task t, we inject samples of the input X_t into a memory buffer with a fixed size M, and the memory will be used to select samples for the purpose of replay when the model is learning on subsequent tasks. We will demonstrate the comparison between different sample selection strategies in Section 4.

3.2. Augmented Acoustic Scenes

Our mutual information optimization relies on the comparisons between different augmented representations of acoustic scenes, which are also called pseudo-labeled samples. In MI estimation, the augmentations of the same input will be regarded as positive pairs, while those of different inputs are taken as negative pairs, such that the dependency between positive pairs will be maximized and that between negative pairs will be minimized [17]. Following the notations in Section 3.1, we will denote Z as the feature representation of the original input X, and Z' as the encoded feature of an augmentation of input X'.

In this work, we simulate the pseudo-labeled samples through different augmentation methods. More specifically, we choose to add Gaussian noise, apply band-stop filtering, or invert along the time axis to perform multiple types of augmentations.

3.3. Modeling and Memory Selection via Mutual Information Optimization

By employing an arbitrary model architecture of a feature extractor followed by a classifier, we show that our mutual information optimization can be applied to both modules of the model. We would like the feature extractor Φ to learn task-agnostic knowledge to produce effective latent representations of the input audio, while the classifier Θ to learn task-specific knowledge to map the learned representations to specific acoustic scene classes.

Feature extractor part. To let the feature extractor learn task-agnostic knowledge, we need to guarantee that the encoded representations preserve sufficient information from the original inputs regardless of their classes. Therefore, we will want to maximize the MI between X and Z, so that Z will preserve generic information from X in the modeling of the feature extractor. To better estimate the MI, we have augmented acoustic scenes X' and its corresponding latent representation Z' described in Section 3.2. Assuming that Z and Z' are conditional independent given data X, we have

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X) \\ &= H(Z) - H(Z|X,Z') \\ &\geq H(Z) - H(Z|Z') = I(Z;Z') \end{aligned} \tag{1}$$

where $I(\cdot,\cdot)$ indicates the mutual information between two variables and $H(\cdot)$ denotes the Shannon entropy or conditional entropy given another random variable. We use the property of conditional independence from Line 1 to Line 2, and the definition of conditional entropy from Line 2 to Line 3. Therefore, we have $I(X;Z) \geq I(Z;Z')$ from Eq. 1, which means that maximizing the MI between Z and Z' is equivalent to maximizing the lower bound of the MI between input X and the encoded features Z.

Taking a further step, as from [17], the mutual information between Z and its augmentation Z' can be estimated through the InfoNCE (NCE stands for noise-contrastive estimation) loss as the lower bound, i.e.,

$$I(Z,Z') \geq \log N + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\log\frac{(f(z_i,z_i')/\tau)}{\sum_{j=1}^{N}(f(z_i,z_j')/\tau)}}_{\triangleq\mathcal{L}_{\text{NCE}(Z,Z')}} \tag{2}$$

where z_i is the representation of an individual sample in the batch. z_i and z_i' are regarded as positive sample pairs since they originated from the same sample x_i, and all other z_j's in the batch where $i \neq j$ are regarded as negative pairs. f denotes the exponential of similarity function and τ is the temperature. N is the batch size of the samples, and when it becomes larger, $I(Z, Z')$ will close its gap to the lower bound. Therefore, $I(Z, Z')$ is lower bounded by the InfoNCE loss and we will use the second term on the right-hand side as the approximation of MI in our implementations. Overall, in addition to the task supervision loss, we add the MI estimation to the objective function to train the model as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(\Theta(\Phi(X)), Y) + \lambda \mathcal{L}_{\text{NCE}}(Z, Z') \tag{3}$$

where $\mathcal{L}_{\text{CE}}(\cdot, \cdot)$ denotes the cross entropy loss between the predicted and the ground-truth logits, and λ is the hyperparameter.

Classifier part. The classifier takes the latent representation Z as input and predicts the logits Y. In contrast to the feature extractor where the primary goal is to capture task-agnostic features, since we would like the classifier to learn task-specific knowledge, we need to wisely select the samples from the memory, such that they can not only bring extra information but also make sure the new information can be learned by the model. Two criteria, *surprise* and *learnability*, can be formalized by measuring the predictive distribution of the new sample with respect to those in the memory to decide its usefulness [29]. In our work, we measure the information carried by memory samples by estimating the MI between the encoded representation Z and the predicted logits Y. In this case, it differs from the original InfoNCE loss with additional information Y. Since we would like to train the classifier to learn task-specific knowledge given the label information, we use Y additionally to determine positive and negative sample pairs. If we further incorporate the label information Y into Eq. 2, for each sample feature z_i, we consider z_k as positively paired sample for all $y_k = y_i$. In other words, samples with the same predicted labels with respect to sample i will be constructed as positive pairs. Meanwhile, all pairs of the original and augmented representations z_k and z'_k, along with the augmented representation z'_i, are regarded as positive pairs with respect to z_i, while others are taken as negative pairs, i.e.,

$$\begin{aligned}
&\mathcal{L}_{\text{NCE}}(Z, \{Z'\}, Y) \\
&= \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{\sum_{k=1}^{N} \mathbb{1}(y_k = y_i)} \sum_{y_k = y_i} (\sum_{\hat{z_i} \in \mathcal{S}_{z_i}} \log(f(z_i, \hat{z_i})/\tau) \\
&- \log \sum_{j=1}^{N} (\sum_{\hat{z_j} \in \mathcal{S}_{z_j}} f(z_j, \hat{z_j})/\tau))]
\end{aligned} \tag{4}$$

where $\mathcal{S}_{z_i}$ indicates the set of all of the original and augmented views of the sample z_k that has the same label with z_i, along with the augmented view of itself z'_i. The $\mathbb{1}(\cdot)$ function returns 1 if the condition is true and 0 otherwise. Note that here we use the notations of sets, $\{Z'\}$, to indicate that more than one type of augmentation techniques can be considered.

When we are selecting from the memory at task t, we would like to select the samples that are both representative and informative. It not only needs to carry new information to the existing model, but also should have a high learnability instead of becoming an outlier. To achieve this, we assign a score to the samples and select the samples with the highest scores as

$$\begin{aligned}
\text{score}_t(Y, Z) = &-\mathcal{L}_{\text{NCE}}(Z_{t-1}, \{Z'_{t-1}\}, Y_{t-1}) \\
&+ \mathcal{L}_{\text{NCE}}(Z_t, \{Z'_t\}, Y_t)
\end{aligned} \tag{5}$$

The first term on the right-hand side of Eq. 5 indicates that we would like to be more likely to select samples that minimize the mutual information between Z and Z', given the class logits by the previous model at task $t - 1$. In other words, the samples that are more surprising to the model are favored. Similarly, the second term indicates that the samples with higher learnability are more probably to be sampled, since they maximize the MI between Z and Z' given Y by the current model, which aligns with our objective function. In this way, we will be more likely to select samples that are both representative and informative, so that the model can effectively recall past knowledge and learn from new information at the same time.

Method	Memory Size	Acc ↑	BWT ↑	FWT ↑
fine-tune	-	20.4	-56.0	0.0
Random	0.2k	42.8	-28.5	49.8
	0.5k	49.8	-27.8	54.3
	1k	52.6	-27.0	59.2
Herding [24]	0.2k	51.6	-26.9	56.0
	0.5k	54.3	-26.3	63.3
	1k	56.2	-24.8	65.2
GSS [25]	0.2k	51.9	-25.3	56.5
	0.5k	54.6	-25.8	62.9
	1k	56.1	-24.6	63.7
Uncertainty [27]	0.2k	55.9	-24.5	63.8
	0.5k	57.6	-23.7	67.5
	1k	58.9	-22.8	69.0
MIO (Ours)	0.2k	58.0	-23.5	64.7
	0.5k	60.7	-22.9	69.1
	1k	**64.1**	**-22.5**	**74.8**

Table 1. Results for continual learning on TAU Urban Acoustic Scenes with different memory selection methods and sizes.

4. EXPERIMENTS

4.1. Datasets

We use the TAU Urban Acoustic Scenes 2022 [32] and Environmental Sound Classification (ESC)-50 [33] as our datasets. TAU Urban Acoustic Scenes consists of 10 classes of acoustic scenes in total, with around 1,000 samples for each class. Its acoustic scene classes are mainly about transportation and city noises. ESC-50 is a smaller dataset that is made up of 5-second-long recordings in 50 semantical classes, with 40 samples for each class. This dataset mainly covers the sounds from animals, humans and daily activities, etc. The diversity of these datasets helps us validate the effectiveness of our continual learning methods under different settings. For the task splitting, as described in Section 3.1, we split the 10 or 50 classes in the two datasets into 5 sequential tasks $\{X_t, Y_t^{\text{gt}}\}_{t=1}^{5}$. Each task contains 2 or 10 different classes respectively.

4.2. Baselines

We evaluate the continual learning performance of acoustic scene classification with multiple baseline approaches together with our proposed method as in Tables 1 and 2. *Fine-tune* means the offline training without any continual learning approaches performed, which is the lower bound of our performance. *Random* is to randomly select samples from the memory with an equal probability. *Herding* indicates herding the embeddings of samples and selects those who are closest to the center of their corresponding class [24]. *GSS* refers to gradient-based sample selection, which aims to maximize the diversity of the gradients of the samples in the memory buffer [25]. *Uncertainty* calculates the uncertainty score of each sample based on the prototypes from the herding method, and selects

Method	Memory Size	Acc ↑	BWT ↑	FWT ↑
fine-tune	-	19.1	-58.7	0.0
Random	0.2k 0.5k 1k	22.5 24.6 26.2	-52.5 -49.7 -47.6	26.6 27.3 29.7
Herding [24]	0.2k 0.5k 1k	47.5 49.3 50.8	-30.8 -28.7 -27.9	49.3 50.6 52.2
GSS [25]	0.2k 0.5k 1k	48.8 49.6 50.3	-30.3 -29.3 -28.2	49.8 50.8 51.9
Uncertainty [27]	0.2k 0.5k 1k	50.9 51.8 52.9	-28.9 -27.6 -27.1	51.6 53.1 53.9
MIO (Ours)	0.2k 0.5k 1k	52.1 53.7 **55.3**	-28.5 -27.4 **-26.5**	53.4 55.9 **57.3**

Table 2. Results for continual learning on Environmental Sound Classification-50 dataset with different memory selection methods and memory sizes.

the samples that the model is less confident of [27]. We will compare the performance of our proposed method with these baselines as different memory selection mechanisms in replay-based continual learning methods.

4.3. Experimental Setups

For our feature extractor, we use a Temporal Convolutional Network (TCN) [34] as the feature extractor and a linear layer as the classifier. The feature extractor takes in the log-Mel spectrogram of the audio input, computed by a Hanning window with the window length of 25ms and the hop length of 10ms. The latent representation Z is represented as 100-dim embedding vectors, which are used to compute the scoring function to sample from the memory. We train the model for 50 epochs for each task with an Adam optimizer and an initial learning rate of 0.0005. We use the temperature $\tau = 1.0$ for all the experiments.

4.4. Evaluation Metrics

Aside from average classification accuracy (Acc), we use backward transfer (BWT) and forward transfer (FWT) as the evaluation metrics [35] to show that our method helps not only learn task-agnostic knowledge, but also preserve the task-specific knowledge. BWT measures the influence of learning task t on the accuracies of all previous tasks $i < t$. The calculation of BWT at task t is defined as

$$\text{BWT}_t = \frac{2}{t(t-1)} \sum_{i=2}^{t} \sum_{j=1}^{i-1} (a_{t,i} - a_{i,i}), t \in \{2, \cdots, T\}$$

where $a_{i,j}$ denotes the accuracy of task j after learning task i. On the contrary, FWT measures the generalizability of the model by computing the influence of learning task t on the accuracies of future tasks. From [10], we have

$$\text{FWT}_t = \frac{1}{t-1} \sum_{i=2}^{t} (a_{i-1,i} - \bar{a}_i), t \in \{2, \cdots, T\}$$

where $\bar{a}_i$ indicates the test accuracy at task i with random initialization. Overall, a higher BWT score means a smaller forgetting effect of the model on past task-specific knowledge, while a higher FWT score means a higher generalizability of the model on task-agnostic knowledge to benefit unseen tasks.

4.5. Results and Discussion

The experimental results on the TAU Urban Acoustic Scene dataset are shown in Table 1. The row of *fine-tune* suffers from catastrophic forgetting, and its accuracy is close to a random guess since we have 5 tasks in total. It is the lower bound of our continual learning performances. For the rest of the rows, we can observe that our proposed mutual information optimization (MIO) method achieves the highest score of Acc, BWT, FWT than other memory selection methods, which indicates that it can not only retain the task-specific knowledge in the past, but generalize the task-agnostic knowledge to future unseen classes as well. Another observation is that our method benefits more from a larger memory size, with a higher performance gain from a smaller memory size to a larger one. This intuition aligns well with the property of the estimates of mutual information in Eq. 2, where the estimated MI approaches its lower bound when the number of samples N becomes larger.

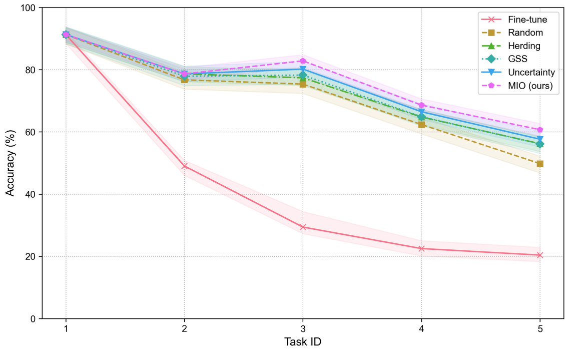

Fig. 1. Average Acc (%) over tasks in sequential order for different methods. The accuracies are calculated on the test sets of the seen tasks so far.

Table 2 presents the results on the ESC-50 dataset with the same set of evaluation metrics. Overall, we can observe a similar tendency with the results in Table 1, except that the accuracies become lower because there are 50 classes in total and less number of samples per class. We also plot the change of the accuracies over the sequential tasks in Figure 1. From the figure, we can observe that our method has the least forgetting effect with a smaller decrease in accuracy. In contrast, fine-tuning has the largest drop with accuracy close to $\frac{1}{t}$, where t is the number of tasks that the model has experienced. It is noteworthy that from task 2 to task 3, some continual learning methods have an increased accuracy instead of a decreasing one. We speculate that this phenomenon is due to the shared properties of the acoustic scene classes that these tasks consist of.

5. CONCLUSION

This paper presents a replay-based continual learning approach in the acoustic scene classification task with mutual information estimation. We propose to optimize different levels of the model to learn task-agnostic and task-specific knowledge from the perspective of mutual information, and select samples from the memory buffer that are both representative and informative. We demonstrate that our proposed method outperforms other continual learning algorithms by both a lower forgetting effect and higher generalizability.

Acknowledgement: We would like to thank Shinji Watanabe for his valuable comments.

6. REFERENCES

[1] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Assessment of human and machine performance in acoustic scene classification: Dcase 2016 case study," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 319–323.

[3] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *Proc. Interspeech*, 2016.

[4] Karol J Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2015, pp. 1–6.

[5] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu, "A comprehensive survey of continual learning: Theory, method and application," *arXiv preprint arXiv:2302.00487*, 2023.

[6] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[7] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[8] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[9] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4548–4557.

[10] David Lopez-Paz and Marc'Aurelio Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny, "Efficient lifelong learning with a-gem," in *International Conference on Learning Representations*, 2019.

[12] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim, "Continual learning with deep generative replay," *Advances in neural information processing systems*, vol. 30, 2017.

[13] Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi, "Rainbow memory: Continual learning with a memory of diverse samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8218–8227.

[14] Richard Kurle, Botond Cseke, Alexej Klushyn, Patrick Van Der Smagt, and Stephan Günnemann, "Continual learning with bayesian neural networks for non-stationary data," in *International Conference on Learning Representations*, 2020.

[15] Michele Valenti, Stefano Squartini, Aleksandr Diment, Giambattista Parascandolo, and Tuomas Virtanen, "A convolutional neural network approach for acoustic scene classification," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1547–1554.

[16] Kristi Hendrickson, Matthew Walenski, Margaret Friend, and Tracy Love, "The organization of words and environmental sounds in memory," *Neuropsychologia*, vol. 69, pp. 67–76, 2015.

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[18] Philip Bachman, R Devon Hjelm, and William Buchwalter, "Learning representations by maximizing mutual information across views," *Advances in neural information processing systems*, vol. 32, 2019.

[19] Olivier Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International conference on machine learning*. PMLR, 2020, pp. 4182–4192.

[20] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Self-supervised representation learning with relative predictive coding," in *International Conference on Learning Representations*, 2021.

[21] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed, "Information maximization for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2445–2457, 2020.

[22] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[23] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," in *International Conference on Learning Representations*, 2018.

[24] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[25] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio, "Gradient based sample selection for online continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[26] Muqiao Yang, Ian Lane, and Shinji Watanabe, "Online continual learning of end-to-end speech recognition models," *Proc. Interspeech*, 2022.

[27] Yang Xiao, Xubo Liu, James King, Arshdeep Singh, Eng Siong Chng, Mark D Plumbley, and Wenwu Wang, "Continual learning for on-device environmental sound classification," *arXiv preprint arXiv:2207.07429*, 2022.

[28] Umberto Cappellazzo, Daniele Falavigna, and Alessio Brutti, "Exploring the joint use of rehearsal and knowledge distillation in continual learning for spoken language understanding," *arXiv preprint arXiv:2211.08161*, 2022.

[29] Shengyang Sun, Daniele Calandriello, Huiyi Hu, Ang Li, and Michalis Titsias, "Information-theoretic online memory selection for continual learning," in *International Conference on Learning Representations*, 2022.

[30] Yiduo Guo, Bing Liu, and Dongyan Zhao, "Online continual learning through mutual information maximization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8109–8126.

[31] Yujun Shi, Li Yuan, Yunpeng Chen, and Jiashi Feng, "Continual learning via bit-level information preserving," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 16674–16683.

[32] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," *arXiv preprint arXiv:2005.14623*, 2020.

[33] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018, ACM Press.

[34] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[35] Natalia Daz-Rodrguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni, "Don't forget, there is more than forgetting: new metrics for continual learning," *arXiv preprint arXiv:1810.13166*, 2018.