Proceedings Paper:

Ye, Fei and Bors, Adrian Gheorghe orcid.org/0000-0001-7838-0021 (2023) Continual Variational Autoencoder via Continual Generative Knowledge Distillation. In: AAAI Conference on Artificial Intelligence. AAAI Press

Continual Variational Autoencoder via Continual Generative Knowledge Distillation

Fei Ye and Adrian G. Bors

Department of Computer Science, University of York, York YO10 5GH, UK
fy689@york.ac.uk, adrian.bors@york.ac.uk

Abstract

Humans and other living beings have the ability of short and long-term memorization during their entire lifespan. However, most existing Continual Learning (CL) methods can only account for short-term information when training on infinite streams of data. In this paper, we develop a new unsupervised continual learning framework consisting of two memory systems using Variational Autoencoders (VAEs). We develop a Short-Term Memory (STM), and a parameterised scalable memory implemented by a Teacher model aiming to preserve the long-term information. To incrementally enrich the Teacher's knowledge during training, we propose the Knowledge Incremental Assimilation Mechanism (KIAM), which evaluates the knowledge similarity between the STM and the already accumulated information as signals to expand the Teacher's capacity. Then we train a VAE as a Student module and propose a new Knowledge Distillation (KD) approach that gradually transfers generative knowledge from the Teacher to the Student module. To ensure the quality and diversity of knowledge in KD, we propose a new expert pruning approach that selectively removes the Teacher's redundant parameters, associated with unnecessary experts which have learnt overlapping information with other experts. This mechanism further reduces the complexity of the Teacher's module while ensuring the diversity of knowledge for the KD procedure. We show theoretically and empirically that the proposed framework can train a statistically diversified Teacher module for continual VAE learning which is applicable to learning infinite data streams.

Introduction

The continuous acquisition and learning of new concepts from a dynamically evolving environment is a fundamental function for an artificial intelligent system. Modern deep learning models have already outperformed humans in learning certain single tasks (LeCun, Bengio, and Hinton 2015), but would suffer dramatic performance degeneration when attempting to learn a sequence of different data domains. The phenomenon of such a performance degeneration is referred in AI as "catastrophic forgetting".

Existing work on continual learning mainly focuses on the classification task and requires knowledge of the task information during training. In this work, we focus on the lifelong generative modelling (Ramapuram, Gregorova, and Kalousis 2017; Ye and Bors 2020a) and consider a more sophisticated learning scenario, called the Task-Free Continual Learning (TFCL) (Aljundi, Kelchtermans, and Tuytelaars 2019), which would not require the task identity during the training and testing. One popular way to reduce forgetting in TFCL is the replay-based approach, which either manages a fixed-capacity memory buffer (De Lange and Tuytelaars 2021; Jin et al. 2021) or trains a generator (Shin et al. 2017) which then can be used to reproduce the data through a generative replay mechanism. The former approach requires a suitable sample selection criterion that selectively stores incoming samples into the memory to relieve forgetting (Aljundi et al. 2019b; De Lange and Tuytelaars 2021) while the latter aims to train a powerful generator that produces high-quality generative replay samples, which are statistically consistent with the training sets (Ye and Bors 2020a). However, these approaches rely on a single memory system, which is not scalable for learning infinite data streams due to their limited memory capacity, while also requiring frequent retraining (Ye and Bors 2020a).

Recently, the Dynamic Expansion Model (DEM) (Lee et al. 2020; Rao et al. 2019) has shown promising results for TFCL and can potentially be applied to infinite data streams. The expansion criterion in DEM plays an important role in balancing the complexity of the model and the generalization performance. Existing DEM-based methods implement the expansion criterion by considering the sample log-likelihood evaluation (Rao et al. 2019) and Dirichlet processes (Lee et al. 2020), but they cannot guarantee an optimal trade-off between the network architecture and performance (See Theorem 2). They also have a multi-head structure that requires performing the component selection at the testing phase. In addition, they cannot model correlations between different data domains in a single latent space leading to additional computational resources. At the same time, its application is rather limited to specific tasks such as cross-domain image interpolation (Oring, Yakhini, and Hel-Or 2021; Ye and Bors 2020a, 2022a,c,b).

In this paper, we study lifelong generative modelling under TFCL, which aims to learn a model capable of generating images for all previously learnt data domains without forgetting and without accessing any task information. To address this issue, we propose a new knowledge distillation framework that aims to train a compact model as a parameterised memory (Teacher module) together with a short-term memory holder model to enable training a continual Variational Autoencoder (VAE) as a Student, under TFCL. To achieve these goals, we propose the Knowledge Incremental Assimilation Mechanism (KIAM) that evaluates the probability distance between the current memory buffer and the already accumulated knowledge by the Teacher module as signals to increase the Teacher's memorization capacity. This mechanism can ensure the knowledge diversity and prevent forgetfulness in the Teacher module. Moreover, we propose a new data-free KD approach, namely the Continual Generative Knowledge Distillation (CGKD), which transfers generative knowledge from the Teacher to the Student without accessing real data samples. To further ensure the diversity of knowledge for KD, we propose a new expert pruning approach that selectively removes redundant experts from the Teacher module, aiming to reduce its complexity. Another contribution consists in the derivation of the upper bound for the negative sample log-likelihood, which provides new insights into the forgetting behaviour and theoretical guarantees for the proposed framework.

In summary, our contributions are as follows: 1) We propose a new KD framework for lifelong generative modelling under the challenging TFCL learning setting; 2) We propose the Knowledge Incremental Assimilation Mechanism (KIAM) for the Teacher module, which enriches its knowledge incrementally while also ensuring a minimal architecture; 3) A new KD approach is introduced to transfer generative knowledge from a Teacher to a Student module in an online manner; 4) We propose a new expert pruning approach for KD, which reduces the size of the Teacher module while preserving its knowledge diversity. 5) This is the first study to provide theoretical insights for lifelong generative modelling without knowing any task information.

Supplementary materials (SM) and source code are available[1].

Related Work

Knowledge Distillation : The transfer of knowledge from a complex Teacher model to a lightweight Student network, called Knowledge Distillation (KD), has recently been studied (Heo et al. 2019; Hinton, Vinyals, and Dean 2014). Most KD approaches use a fixed Teacher model trained on the target dataset. Then a classifier is trained as a Student using Teacher's generated predictions. The Teacher model can also be implemented by an ensemble of networks framework where the Student learns multi-mode knowledge from the Teacher to improve its generalization performance (Phuong and Lampert 2019; Nam et al. 2021). Recently, KD has been used for continual learning while relieving forgetting (Li and Hoiem 2017; Zhai et al. 2019). In these approaches, the previously trained classifier is considered as the Teacher while a new classifier is trained as the Student through the KD process, where the Student's predictions are forced to match the Teacher's outputs (Hinton, Vinyals, and Dean 2014; Buzzega et al. 2020). However, these approaches still require task information, which is not suitable for TFCL.

Continual learning : Most existing approaches to CL require access to task information during training. There are three categories of CL approaches : Regularisation-based methods (Hinton, Vinyals, and Dean 2014; Kirkpatrick et al. 2017; Kurle et al. 2020; Li and Hoiem 2017; Nguyen et al. 2018; Polikar et al. 2001; Ren et al. 2017; Ritter, Botev, and Barber 2018), dynamic architectures (Fernando et al. 2017; Golkar, Kagan, and Cho 2019; Hung et al. 2019; Rusu et al. 2016; Wen, Tran, and Ba 2020; Ye and Bors 2023, 2021a, 2022d) and memory-based approaches (Achille et al. 2018; Ramapuram, Gregorova, and Kalousis 2017; Rao et al. 2019; Shin et al. 2017; Sun et al. 2022; Ye and Bors 2020a,b, 2022e; Zhai et al. 2019; Ye and Bors 2021b; Yoon et al. 2022). Because of requiring the task information (Aljundi et al. 2019a) these approaches cannot be applied directly to TFCL. A popular method for TFCL is to have a small memory buffer containing training samples to relieve forgetting (Aljundi et al. 2019a; Aljundi, Kelchtermans, and Tuytelaars 2019). The sample selection criterion for the memory buffer plays a key role in the performance and can rely on the parameters updating gradient (Aljundi et al. 2019b) or on a specific loss function (De Lange and Tuytelaars 2021) during training. Moreover, the stored samples can be edited, aiming to adapt them to learning new tasks while preserving the previously learnt information (Jin et al. 2021). However, these approaches can only be used in a fixed-length data stream learning context.

Lifelong generative modelling : Generative modelling in continual learning aims to train a generative model capable of producing data generations and reconstructions without forgetting (Egorov, Kuzina, and Burnaev 2021; Ramapuram, Gregorova, and Kalousis 2017). The Variational Autoencoder (VAE) (Kingma and Welling 2013) was firstly explored for lifelong generative modelling in (Achille et al. 2018), while the Generative Adversarial Network (GAN) was used in (Ye and Bors 2021b). The Generative Replay Mechanism (GRM) was used to relieve forgetting by reproducing data learnt in the past. More recently, the GRM has been extended to several frameworks, including the Teacher-Student structure (Ramapuram, Gregorova, and Kalousis 2017; Ye and Bors 2022e) and the VAE-GAN hybrid model (Ye and Bors 2020a). However, these frameworks require the task information and cannot be applied in TFCL. On the other hand, the Dynamic Expansion Model (DEM) is a promising approach for generative modelling in TFCL. The first work using DEM to TFCL, proposed in (Rao et al. 2019), was called the Continual Unsupervised Representation Learning (CURL). In CURL an expansion mechanism is introduced to dynamically build new inference models for capturing new concepts during the training. A similar idea was used in the Continual Neural Dirichlet Process Mixture (CN-DPM) (Lee et al. 2020), which dynamically creates VAE-based experts in a mixture system through a Dirichlet process-based expansion mechanism. However, these approaches lead to non-optimal architectures as they ignore accounting for the knowledge diversity when performing model expansion.

[1] https://github.com/dtuzi123/CGKD

Methodology

Problem Definition

Let $D_i^T = \{\mathbf{x}_j^T\}_{j=1}^{N_i^T}$ and $D_i^S = \{\mathbf{x}_j^S\}_{j=1}^{N_i^S}$ be the test and training sets, for the i-th data domain, where N_i^T and N_i^S represent the number of samples for the test and training sets, respectively. In this work, we focus on sequential learning different data domains without accessing the task/domain information. Therefore, we continuously create a data stream $\mathcal{S}$ by including all incoming training sets $\mathcal{S} = \sum_{i=1}^{k}\{S \cup D_i^S\}$. However, at a training step $\mathcal{T}_m$, we only access a small batch of samples $\{\mathbf{x}_{m,j}\}_{j=1}^{b}$ drawn from $\mathcal{S}$, where b is the batch size. We assume that learning the whole data set $\mathcal{S}$ requires a total of t training steps. After a model is done with the training step $\mathcal{T}_t$, we evaluate the performance of the model on all test sets $\{D_1^T, \cdots, D_k^T\}$.

Knowledge Incremental Assimilation Mechanism

Humans can incrementally learn and memorize novel concepts throughout their entire lifespan (Banayeeanzade et al. 2021). Inspired by this, we introduce the Knowledge Incremental Assimilation Mechanism (KIAM) for the Teacher module, which gradually increases the knowledge capacity of the Teacher. Let $\mathcal{M}_i$ represent a short-term memory (STM) updated at $\mathcal{T}_i$ and $\mathbf{A}_i = \{\mathcal{A}_1, \cdots, \mathcal{A}_c\}$ be a Teacher model assumed to have already trained c experts up to the training step ($\mathcal{T}_i$), where each $\mathcal{A}_j$ is implemented by either a GAN (Goodfellow et al. 2014) or a VAE to learn a generator distribution $\mathbb{P}_{\theta_j}$ with trainable parameters θ_j. Detecting when and what new concepts are provided is a real challenge under the TFCL framework, where we do not have task labels. To address this problem, KIAM evaluates probabilistic distances between the current memory and the already learnt information aiming to detect when the data distribution shift would occur, according to :

$$\min_{j=1}^{c-1} D_p(\mathbb{P}_{\theta_j}, \mathbb{P}_{\mathcal{M}_i}) \geq \nu \;, \tag{1}$$

where $\mathbb{P}_{\mathcal{M}_i}$ is the probability representation characterizing the current memory buffer $\mathcal{M}_i$ and $\nu \in [0, 200]$ is a threshold for controlling the number of experts for the Teacher module. We omit the current expert $\mathcal{A}_c$ in Eq. (1) since $\mathcal{A}_c$ is knowledgeable about $\mathcal{M}_i$. $D_p(\cdot,\cdot)$ is used to evaluate the knowledge similarity between the current memory $\mathcal{M}_i$ and each previously learnt expert. We consider the Fréchet Inception Distance (FID) (Heusel et al. 2017) as the $D_p(\cdot,\cdot)$ measure for evaluating the knowledge similarity between two probabilities, because FID is an non-parametric measure which does not require explicit probabilistic representations. If Eq. (1) is satisfied, we freeze $\mathcal{A}_c$ that has already preserved the knowledge of the current memory $\mathcal{M}_i$, while adding a new expert $\mathcal{A}_{c+1}$ for next training step. We also empty the current memory in order for the newly created expert $\mathcal{A}_{c+1}$ to be able to learn non-overlapping probability densities.

Continual Generative Knowledge Distillation

Existing approaches to knowledge distillation usually transfer the category information from a complex Teacher module to a lightweight Student for the classification task. However, these methods are not able to distil the knowledge for generative modelling because previously learnt samples are not available during the continual learning process. In this paper, we introduce a data-free KD approach for lifelong generative modelling. Let us consider a latent variable based generative model $p_\xi(\mathbf{x}, \mathbf{z}) = p_\xi(\mathbf{x} \,|\, \mathbf{z})p(\mathbf{z})$, which is represented by the Student module in our framework, where $\mathbf{x}$ and $\mathbf{z}$ represent the observed and latent variables, respectively. $p_\xi(\mathbf{x} \,|\, \mathbf{z})$ and $p(\mathbf{z})$ are the decoding and prior distributions, respectively. An approach for distilling knowledge when having a Teacher module characterized by a generator distribution $\mathbb{P}_{\theta_j}$, would require to minimise the KL divergence between $\mathbb{P}_{\theta_j}$ and $p_\xi(\mathbf{x})$. However, this is computationally infeasible due to the lack of explicit density function for $\mathbb{P}_{\theta_j}$. Instead, we propose to implement KD by minimizing the cross entropy between $\mathbb{P}_{\theta_j}$ and $p_\xi(\mathbf{x})$:

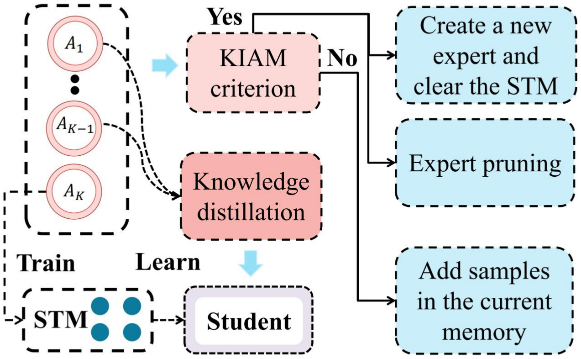

Figure 1: The learning procedure of the proposed framework where we omit the updating of the memory for the sake of simplification. During the training, we optimize the current teacher component and the student module using the adversarial and VAE loss, respectively. We then check the model expansion and perform expert pruning.

$$\mathcal{L}_{KD} = \sum\nolimits_{j=1}^{c-1} \big\{ -\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\theta_j}} [\log p_\xi(\mathbf{x})] \big\} \,. \tag{2}$$

The direct calculation of Eq. (2) is not possible because the marginal log-likelihood $\log p_\xi(\mathbf{x}) = \log \int p_\xi(\mathbf{x} \,|\, \mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$ requires the integration over all variables $\mathbf{z}$ from $p(\mathbf{z})$. In this paper, we introduce the use of a variational distribution $q_\phi(\mathbf{z} \,|\, \mathbf{x})$, parameterised by ϕ, to approximate the true posterior $p_\xi(\mathbf{z} \,|\, \mathbf{x})$, and therefore the marginal log-likelihood $\log p_\xi(\mathbf{x})$ can be estimated by a lower bound (Kingma and Welling 2013). Then, Eq (2) is calculated as :

$$\begin{aligned}\mathcal{L}_{KD} = \sum\nolimits_{j=1}^{c-1} \big\{ &-\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\theta_j}} [\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})} [\log p_\xi\left(\mathbf{x} \,|\, \mathbf{z}\right)] \\ &- KL\left[q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right) \,||\, p\left(\mathbf{z}\right)\right]] \big\} \,.\end{aligned} \tag{3}$$

Together with the KD loss Eq. (3), and using the current memory buffer $\mathcal{M}_i$ for training the model at $\mathcal{T}_i$, we design a unified objective function for training the student module at $\mathcal{T}_i$ as :

$$\begin{aligned}\mathcal{L}_{Stu} = &-\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\mathcal{M}_i}} [\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})} [\log p_\xi\left(\mathbf{x} \,|\, \mathbf{z}\right)] \\ &- KL\left[q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right) \,||\, p\left(\mathbf{z}\right)\right]] + \mathcal{L}_{KD} \,,\end{aligned} \tag{4}$$

where the first term encourages the Student module to learn samples drawn from $\mathcal{M}_i$ and the second term $\mathcal{L}_{KD}$ transfers the Teacher's knowledge to the Student.

Expert Pruning for the KD Procedure

The Teacher's ensembles cannot grow forever and in order to keep the number of parameters in check and architecture compact while preserving the statistical representation diversity of the Teacher, we consider a component expert pruning approach. We find and remove the component with the highest statistical overlap, after defining a measure of knowledge similarity between two expert components.

Suppose that the Teacher module has already trained c experts, $\mathbf{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_c\}$, let $\mathbf{Q} \in \mathcal{R}^{c \times c}$ be a knowledge discrepancy matrix, where each $\mathbf{Q}(a,b)$ represents the discrepancy score between experts $\mathcal{A}_a$ and $\mathcal{A}_b$. Given that the Student module is already knowledgeable about the informational content of all experts, it can be used for identifying whether two experts contain statistically overlapping knowledge. We evaluate $\mathbf{Q}(a,b)$ by calculating the square loss $\| \cdot \|^2$ on the latent variables, inferred by the inference model $q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right)$ of the student :

$$\mathcal{L}_{ks}(\mathcal{A}_a, \mathcal{A}_b) = \mathbb{E}_{\mathbf{x}_a \sim \mathcal{A}_a, \mathbf{x}_b \sim \mathcal{A}_b} \| f_\phi(\mathbf{x}_a) - f_\phi(\mathbf{x}_b) \|^2 \,, \tag{5}$$

where $\mathbf{z}_a$ and $\mathbf{z}_b$ are latent variables, returned by the inference model $f_\phi := q_\phi\left(\mathbf{z} \,|\, \mathbf{x}\right)$ that receives the data samples $\mathbf{x}_a$ and $\mathbf{x}_b$ generated by $\mathcal{A}_a$ and $\mathcal{A}_b$, respectively. Eq. (5) is computational efficient since is evaluated on the low-dimensional latent space. If the two experts, $\mathcal{A}_a$ and $\mathcal{A}_b$ share significant information, their corresponding latent variables $\mathbf{z}_a$ and $\mathbf{z}_b$ tend to be similar, resulting in a small $\mathcal{L}_{ks}(\mathcal{A}_a, \mathcal{A}_b)$ in Eq. (5). Other criteria can be adapted for the evaluation in Eq. (5) (See details in **Appendix-I.7** from SM[1]). Once the discrepancy matrix $\mathbf{Q}$ is evaluated, we identify a pair of experts containing overlapping information by searching for the minimal discrepancy score in $\mathbf{Q}$:

$$\{a^\star, b^\star\} = \arg \min_{\{a,b\}=1}^{c} \left\{ \mathbf{Q}(a,b) \right\}, \tag{6}$$

where $a^\star$ and $b^\star$ are the indices of the selected experts. We then evaluate the discrepancy score between each other expert from the Teacher's ensemble and either $\mathcal{A}_{a^\star}$ or $\mathcal{A}_{b^\star}$:

$$\begin{aligned} c^\star = \arg\min \Big\{ & \sum\nolimits_{\substack{j=1 \\ j \neq \{a^\star, b^\star\}}}^{c-1} \left\{ \mathbf{Q}(a^\star, j) \right\}, \\ & \sum\nolimits_{\substack{j=1 \\ j \neq \{a^\star, b^\star\}}}^{c-1} \left\{ \mathbf{Q}(b^\star, j) \right\} \Big\}, \end{aligned} \tag{7}$$

where $c^\star$ represents the index of the selected expert to be removed from the Teacher module. The main goal for Eq. (7) is that we keep the expert with the largest discrepancy scores in the Teacher module to ensure the knowledge diversity between Teacher's experts. Then we continue with identifying other redundant experts from the Teacher module using Eq. (6) and Eq. (7) and designate them for removal. We iteratively remove experts from the Teacher module until the number of experts is equal to a predefined $n \in [3, 10]$. Another approach for defining the number of Teacher's experts is explored in **Appendix-G1** from SM[1].

Implementation

Objective functions. We have two approaches, implementing each expert of the Teacher by using either a VAE or a GAN. For the GAN-based Teacher model, we have a discriminator network D_ϵ parameterized by ϵ and a generator G_{θ_j} parameterized by θ_j. We consider the WGAN objective function (Gulrajani et al. 2017) for training the j-th expert $\mathcal{A}_j = \{D_\epsilon, G_{\theta_j}\}$ at $\mathcal{T}_i$:

$$\begin{aligned} \min_{\mathbb{P}_{\theta_j}} \max_{D_\epsilon \in \Theta} \Big\{ & \mathbb{E}_{\mathbf{x}_i \sim \mathbb{P}_{\mathcal{M}_i}} \left[D_\epsilon\left(\mathbf{x}_i\right) \right] - \mathbb{E}_{\mathbf{x}' \sim \mathbb{P}_{\theta_j}} \left[D_\epsilon\left(\mathbf{x}'\right) \right] \\ & + \gamma \mathbb{E}_{\widehat{\mathbf{x}}' \sim \mathbb{P}_{\widehat{\mathbf{x}}'}} \left[\left(\| \nabla_{\widehat{\mathbf{x}}'} D_\epsilon\left(\widehat{\mathbf{x}}'\right) \|_2 - 1 \right)^2 \right] \Big\}, \end{aligned} \tag{8}$$

where γ is a hyperparameter and the last term is used to ensure the discriminator's Lipschitz constraint, (Gulrajani et al. 2017). $\mathbb{P}_{\widehat{\mathbf{x}}'}$ is defined by data sampled equally from the distribution of the buffer memory $\mathbb{P}_{\mathcal{M}_i}$, representing the incoming data, and from the generator distribution $\mathbb{P}_{\theta_j}$ (Gulrajani et al. 2017). It should be noted that we only need a single discriminator during the training to reduce the whole model's size since only generators are used for knowledge distillation. For the VAE-based Teacher model, we introduce two neural networks to model the encoding $q_\eta(\mathbf{z} \,|\, \mathbf{x})$ and decoding distribution $p_{\theta_j}(\mathbf{x} \,|\, \mathbf{z})$, respectively. The VAE loss for training the j-th expert $\mathcal{A}_j = \{p_{\theta_j}(\mathbf{x} \,|\, \mathbf{z}), q_\eta(\mathbf{z} \,|\, \mathbf{x})\}$ at $\mathcal{T}_i$ is defined as :

$$\begin{aligned} \mathcal{L}_{ELBO}(\mathbf{x}; \mathcal{A}_j) = & \mathbb{E}_{q_\phi(\mathbf{z} \,|\, \mathbf{x})} \left[\log p_{\theta_j}\left(\mathbf{x} \,|\, \mathbf{z}\right) \right] \\ & - KL\left[q_\eta\left(\mathbf{z} \,|\, \mathbf{x}\right) \,||\, p\left(\mathbf{z}\right) \right] . \end{aligned} \tag{9}$$

Eq. (9) represents the Evidence Lower Bound (ELBO) to the marginal log-likelihood. Similar to GAN-based experts, we only need a single encoding distribution $q_\eta(\mathbf{z} \,|\, \mathbf{x})$ during the training. The learning process is illustrated in Fig. 1.

Algorithm. We summarize the training algorithm in five steps : (1) **(Updating the memory buffer $\mathcal{M}_i$).** We update $\mathcal{M}_i$ at $\mathcal{T}_i$ by adding a new batch of samples $\{\mathbf{x}_{i,j}\}_i^b$ drawn from the data stream $\mathcal{S}$ into its buffer if the memory is not full $|\mathcal{M}_i| < |\mathcal{M}|_{max}$, otherwise, we remove the earliest batch of samples included in $\mathcal{M}_i$ and add $\{\mathbf{x}_{i,j}\}_{k=1}^b$; (2) **(Teacher learning).** If the Teacher has only a single expert at the initial training phase, we automatically build a new expert $\mathcal{A}_2$ at the training step $\mathcal{T}_{100}$ while freezing $\mathcal{A}_1$. We train the newly added expert on $\mathcal{M}_i$ using either Eq. (8) or (9); (3) **(Checking the expansion).** To avoid the frequent evaluation, we check the expansion when the memory is full $|\mathcal{M}_i| = |\mathcal{M}|_{max}$. When the expansion criterion is satisfied Eq. (1), we add a new expert $\mathcal{A}_{c+1}$ to the Teacher module while cleaning up the memory $\mathcal{M}_i$; (4) **(Expert pruning for KD).** We remove the non-essential experts from the Teacher module using the proposed expert pruning approach until the number of experts in $\mathbf{A}$ matches n; (5) **(Student learning).** We distill the data generated by the Teacher to the Student while simultaneously learning the information from the current memory $\mathcal{M}_i$ using Eq. (4). Then we return to **Step 1** for the next training step $\mathcal{T}_{i+1}$.

Theoretical Framework

In this section, we extend the theoretical framework (Ye and Bors 2022f,d) got analyzing the model's forgetting behaviour and defining theoretical guarantees for the proposed approach. We start with providing necessary notations and definitions as follows.

Preliminary

Notations. Let $\mathcal{B}$ be a single VAE model which has the decoding $p_\xi(\mathbf{x} \,|\, \mathbf{z})$ and encoding distributions $q_\phi(\mathbf{z} \,|\, \mathbf{x})$, respectively. Let h be a hypothesis function in the space of hypotheses $\{h \in \mathcal{H} \mid \mathcal{H} : \mathcal{X} \to \mathcal{X}\}$ where $\mathcal{X} \in \mathcal{R}^d$ is the data space with d dimensions. We implement $h \in \mathcal{H}$ by the encoding-decoding process of $\mathcal{B}$ evaluated on the error function $\mathcal{L}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ which is bounded, $\forall (\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, \mathcal{L}(\mathbf{x}, \mathbf{x}') \leq C$ for some $C > 0$. In our setting, the error function $\mathcal{L}(\mathbf{x}, \mathbf{x}')$ is implemented as the square-loss function $\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2, \mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

Definition 1 ***(Model risk.)*** *For a given memory distribution $\mathbb{P}_{\mathcal{M}_i}$, a risk for $\mathcal{B}$ on $\mathbb{P}_{\mathcal{M}_i}$ is defined as $\mathcal{E}_{\mathbb{P}_{\mathcal{M}_i}}(h, f_{\mathbb{P}_{\mathcal{M}_i}}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathcal{M}_i}} \mathcal{L}(h(\mathbf{x}), f_{\mathbb{P}_{\mathcal{M}_i}}(\mathbf{x}))$, where $f_{\mathbb{P}_{\mathcal{M}_i}} \in \mathcal{H}$ is an identity function.*

Following from (Ye and Bors 2022f,d), we define the discrepancy distance for measuring the similarity between two distributions.

Definition 2 ***(Discrepancy distance.)*** *Let $\mathbb{P}_{D_j^T}$ be a probability distribution for D_j^T over $\mathcal{X}$. The discrepancy distance on two distributions $\mathbb{P}_{D_i^T}$ and $\mathbb{P}_{D_j^T}$, is defined as:*

$$\mathcal{L}_{\text{disc}}\left(\mathbb{P}_{D_i^T}, \mathbb{P}_{D_j^T}\right) = \sup_{(h,h') \in \mathcal{H}} \Big| \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{D_i^T}} \left[\mathcal{L}\left(h'\left(\mathbf{x}\right), h\left(\mathbf{x}\right)\right)\right] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{D_j^T}} \left[\mathcal{L}\left(h'\left(\mathbf{x}\right), h\left(\mathbf{x}\right)\right)\right] \Big| . \quad (10)$$

Definition 3 ***(Empirical discrepancy distance.)*** *In practice, we only have access to finite training sets $\widehat{D}_i^T$ and $\widehat{D}_j^T$ of sample sizes m_i and m_j, respectively. Let $\widehat{\mathbb{P}}_{D_i^T}$ and $\widehat{\mathbb{P}}_{D_j^T}$ denote the empirical distribution for $\widehat{D}_i^T$ and $\widehat{D}_j^T$, respectively. Then we estimate the discrepancy distance with probability $1 - \delta, \delta \in (0, 1)$:*

$$\begin{aligned}\mathcal{L}_{\text{disc}}\big(\mathbb{P}_{D_i^T}, \mathbb{P}_{D_j^T}\big) \leq \mathcal{L}_{\text{disc}}\big(\widehat{\mathbb{P}}_{D_i^T}, \widehat{\mathbb{P}}_{D_j^T}\big) + \\ 8\big(\text{Re}_{\widehat{D}_i^T}\left(\mathcal{H}\right) + \text{Re}_{\widehat{D}_i^T}\left(\mathcal{H}\right)\big) \\ + 3M\left(\sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{2m_i}} + \sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{2m_j}}\right),\end{aligned} \quad (11)$$

where $M > 0$ and $\text{Re}_{U_\mathbb{P}}$ is the Rademacher complexity (See ***Appendix-A*** *from SM[1]). In the following we use $\widehat{\mathcal{L}}_{\text{disc}}(\cdot)$ to represent the right-hand side (RHS) of Eq.* (11).

Forgetting Analysis When Considering A Single VAE Model

The ELBO which is used as the objective function for the VAE model, can also be used as a criterion for the performance evaluation (Burda, Grosse, and Salakhutdinov 2015; Domke and Sheldon 2018; Kingma and Welling 2013). In this section, we analyze the forgetting behaviour of a single VAE model by deriving the upper bound to the negative ELBO under TFCL.

Theorem 1 *For a given data stream $\mathcal{S}$, let $\mathbb{P}_{\mathbf{x}'(1:i)}$ be a probability distribution for all previously seen i data batches $\{\{\mathbf{x}_{1,j}\}_{j=1}^b, \cdots, \{\mathbf{x}_{i,j}\}_{j=1}^b\} \in \mathcal{S}$ at $\mathcal{T}_i$. Let the decoder of a single VAE model be a Gaussian distribution with a diagonal covariance matrix (the diagonal element is $1/\sqrt{2}$). We then derive an upper bound to the negative ELBO at $\mathcal{T}_i$, as :*

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_{\mathbf{x}'(1:i)}} \Big[- \mathcal{L}_{ELBO}\big(\mathbf{x}'(1:i); h\big)\Big] \leq \; \mathcal{E}_A\big(\mathbb{P}_{\mathbf{x}'(1:i)}, \mathbb{P}_{\mathcal{M}_i}\big) \\ + \mathbb{E}_{\mathbb{P}_{\mathcal{M}_i}} \Big[- \mathcal{L}_{ELBO}\big(\mathbf{x}_{\mathcal{M}_i}; h\big)\Big] + \big|KL_1 - KL_2\big| ,\end{aligned} \quad (12)$$

where $\mathbf{x}'(1:i)$ and $\mathbf{x}_{\mathcal{M}_i}$ are the latent variables drawn from $\mathbb{P}_{\mathbf{x}'(1:i)}$ and $\mathbb{P}_{\mathcal{M}_i}$, respectively. $\mathcal{E}_A(\mathbb{P}_{\mathbf{x}'(1:i)}, \mathbb{P}_{\mathcal{M}_i})$ is defined as :

$$\begin{aligned}\mathcal{E}_A\big(\mathbb{P}_{\mathbf{x}'(1:i)}, \mathbb{P}_{\mathcal{M}_i}\big) = \widehat{\mathcal{L}}_{\text{disc}}\big(\mathbb{P}_{\mathbf{x}'(1:i)}, \mathbb{P}_{\mathcal{M}_i}\big) + \\ \mathcal{E}_{\mathbb{P}_{\mathbf{x}'(1:i)}}\big(h^*_{\mathbf{x}'(1:i)}, f_{\mathbf{x}'(1:i)}\big) + \mathcal{E}_{\mathbf{x}'(1:i)}\big(h^*_{\mathbf{x}'(1:i)}, h^*_{\mathcal{M}_i}\big) ,\end{aligned} \quad (13)$$

where $h^*_{\mathbf{x}'(1:i)} = \arg\min_{h \in \mathcal{H}} \mathcal{E}_{\mathbb{P}_{\mathbf{x}'(1:i)}}(h, f_{\mathbf{x}'(1:i)})$ and $h^*_{\mathcal{M}_i} = \arg\min_{h \in \mathcal{H}} \mathcal{E}_{\mathbb{P}_{\mathbf{x}_{\mathcal{M}_i}}}(h, f_{\mathcal{M}_i})$ are the optimal hypotheses for $\mathbb{P}_{\mathbf{x}'(1:i)}$ and $\mathbb{P}_{\mathcal{M}_i}$, respectively. KL_1 and KL_2 are defined as :

$$\begin{aligned}KL_1 = \mathbb{E}_{\mathbb{P}_{\mathbf{x}'(1:i)}} KL\big(q_{\phi^i}(\mathbf{z} \,|\, \mathbf{x}'(1:i)) \,||\, p(\mathbf{z})\big) , \\ KL_2 = \mathbb{E}_{\mathbb{P}_{\mathcal{M}_i}} KL\big(q_{\phi^i}(\mathbf{z} \,|\, \mathbf{x}_{\mathcal{M}_i}) \,||\, p(\mathbf{z})\big) .\end{aligned} \quad (14)$$

$q_{\phi^i}(\mathbf{z} \,|\, \cdot)$ is the encoder of a single VAE model trained on $\mathcal{M}_i$ at $\mathcal{T}_i$.

The proof is provided in **Appendix-A** from SM[1]. From Theorem 1 we can derive several observations. First, Eq. (12) can measure the knowledge gain and loss of a single VAE model at each training step $\{\mathcal{T}_i \,|\, i = 1, \cdots, t\}$. Second, the discrepancy distance between all previously seen samples and the memory ($\widehat{\mathcal{L}}_{\text{disc}}(\mathbb{P}_{\mathbf{x}'(1:i)}, \mathbb{P}_{\mathcal{M}_i})$) is crucial for the generalization performance of h to the target distribution $\mathbb{P}_{\mathbf{x}'(1:i)}$. If the discrepancy distance term in Eq. (12) increases, the Left-Hand Side (LHS) of Eq. (12) would also increase, leading to a deterioration in performance. This usually happens in the training step when aiming to learnt many data samples, where a fixed-capacity memory cannot store all previously seen samples and therefore the discrepancy distance term becomes large, resulting in forgetfulness. RHS of Eq. (12) is also an upper bound to the negative sample log-likelihood $\mathbb{E}_{\mathbb{P}_{\mathbf{x}'(1:i)}}[-\log p_\theta(\mathbf{x}'(1:i))]$ (See Lemma 2 in **Appendix-C** from SM[1]).

Theoretical Guarantees

In this section, we extend the theoretical analysis for a single VAE model to the proposed approach of using an ensemble of experts, while also providing theoretical guarantees. The proof is provided in **Appendix-D** from SM[1]. We also extend the proposed theoretical analysis for other models and provide new insights for their forgetting behaviour in **Appendix-F** from SM[1].

Theorem 2 *Suppose that the Teacher module has already trained c experts $\mathbf{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_c\}$ on $\mathcal{M}_i$ at $\mathcal{T}_i$. Let h be*

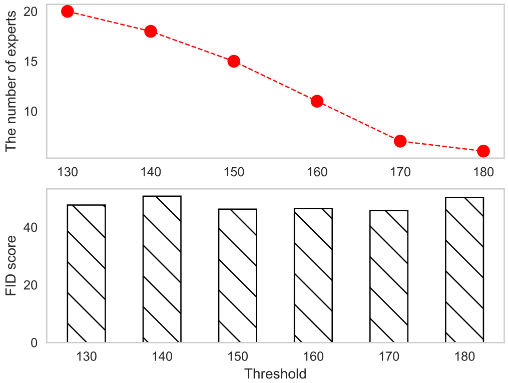

Figure 2: The effect of varying ν.

a Student model which is implemented by a VAE model and we derive an upper bound to the negative ELBO at $\mathcal{T}_i$ as :

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_{\mathbf{x}'(1:i)}}\Big[-\mathcal{L}_{ELBO}\big(\mathbf{x}'(1:i);h\big)\Big] \leq \mathcal{E}_A\big(\mathbb{P}_{\mathbf{x}'(1:i)},\mathbb{P}_{\mathcal{M}_i\otimes\theta_{(1:c)}}\big) \\ +\,\mathbb{E}_{\mathbb{P}_{\mathcal{M}_i\otimes\theta_{(1:c)}}}\Big[-\mathcal{L}_{ELBO}\big(\mathbf{x}'';h\big)\Big] \qquad (15) \\ +\,\big|KL_1 - KL_{\mathcal{M}_i\otimes\theta_{(1:c)}}\big|\,,\end{aligned}$$

where the Kullback-Leibler divergence $KL_{\mathcal{M}_i\otimes\theta_{(1:c)}} = KL(q_{\phi^i}(\mathbf{z}\,|\,\mathbf{x}'')\,||\;\;p(\mathbf{z}))$ and $\mathbf{x}''$ is the latent variable drawn from $\mathbb{P}_{\mathcal{M}_i\otimes\theta_{(1:c)}}$ which is a probability distribution formed by the samples uniformly drawn from $\{\mathbb{P}_{\theta_1},\cdots,\mathbb{P}_{\theta_c},\mathcal{P}_{\mathcal{M}_i}\}$. To compare with a single model (Theorem 1), the proposed Teacher-Student framework can significantly reduce forgetting by increasing the Teacher's knowledge using the KIAM mechanism, while transferring its knowledge to the Student h, according to Eq. (15) in which the discrepancy distance term $\mathcal{E}_A(\mathbb{P}_{\mathbf{x}'(1:i)},\mathbb{P}_{\mathcal{M}_i\otimes\theta_{(1:c)}})$ would be stable if the Teacher gains more knowledge. The study from (Ye and Bors 2022d) also provides a similar bound to Eq. (15). However, the bound in (Ye and Bors 2022d) relies on the task label, which can not be applied in TFCL. Moreover, (Ye and Bors 2022d) does not analyze the knowledge diversity of the Student module, which is one of our major theoretical contributions.

We can also observe that the diversity of the generator distributions $\{\mathbb{P}_{\theta_1},\cdots,\mathbb{P}_{\theta_c}\}$ can reduce the discrepancy distance term by using a minimal number of experts. Inspired by the component diversity analysis (Ye and Bors 2022f), one way to increase the expert diversity is by maximizing the distance between the existing trained experts $\{\mathcal{A}_1,\cdots,\mathcal{A}_{c-1}\}$ and the current expert $\mathcal{A}_c$, expressed as :

$$\mathbb{P}^{\star}_{\theta^m_c} = \arg\max_{\{\mathbb{P}_{\theta^m_c}\,|\,m=i+1,\cdots,t\}}\frac{1}{c-1}\sum\nolimits_{j=1}^{c-1}\left\{D_p(\mathbb{P}_{\theta_j},\mathbb{P}_{\theta^m_c})\right\}, \qquad (16)$$

where $\mathbb{P}^{\star}_{\theta^m_c}$ is an optimal solution and m is the index of the training step. Eq. (16) needs to access all future training steps $\{\mathcal{T}_{i+1},\cdots,\mathcal{T}_t\}$ simultaneously, which is not feasible in continual learning. Instead, the proposed criterion from Eq. (1) can implement the goal expressed in Eq. (16) since it evaluates the probability distance between all previously learnt experts and the current memory. To avoid accessing all future training steps at the same time, Eq. (1) uses a threshold ν to dynamically increase the Teacher's capacity, which acquires new knowledge while preventing forgetting. However, existing dynamic expansion approaches (Lee et al. 2020; Rao et al. 2019) do not fulfil the objective of Eq. (16) as they do not consider the diversity of knowledge learnt by the experts when performing model expansion.

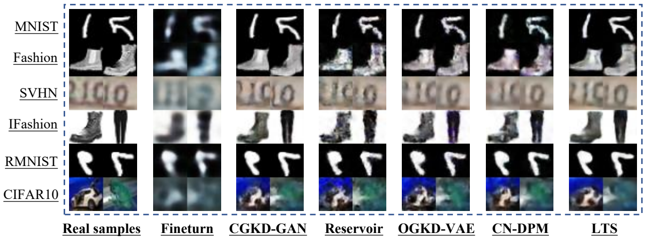

Figure 3: The cross-domain reconstruction results of various models under MSFIRC setting.

Methods	MNIST	SVHN	Fashion	IFashion	RMNIST	CIFAR10	Average	No
finetune	174.1	148.3	237.0	229.1	159.2	216.4	194.0	1
Reservoir	127.2	159.3	213.4	201.6	110.2	113.3	154.2	1
LTS	44.8	62,9	92.9	83.1	41.8	80.3	67.7	1
LGM	104.8	134.3	194.3	168.1	94.8	91.5	131.3	1
CN-DPM	118.7	73.4	120.7	120.3	97.9	97.6	104.8	18
CGKD-GAN	11.6	70.6	101.9	29.9	11.41	68.6	49.0	16
CGKD-VAE	122.9	73.6	109.2	104.3	119.1	86.4	102.6	11
CGKD*-GAN	12.0	74.6	69.8	22.3	11.4	68.5	**43.1**	7
CGKD*-VAE	82.6	82.5	127.0	132.9	88.8	86.3	100.0	7

Table 1: FID for various models under the MSFIRC setting.

Experiments

Settings and Baselines

Setting : We consider a series of six data domains including MNIST (LeCun et al. 1998), SVHN (Netzer et al. 2011), Fashion (Xiao, Rasul, and Vollgraf 2017), IFashion, RMNIST and CIFAR10 (Krizhevsky and Hinton 2009). We create a data stream $\mathcal{S}$ from all training sets of these data domains, namely MSFIRC. We also consider the class-incremental setting where we split each data domain into five parts, and each part consists of images from two different classes (De Lange and Tuytelaars 2021). We then create a data stream $\mathcal{S}$ by combining all parts of the six data domains, namely Class Incremental (CL)-MSFIRC. The batch size and the number of epochs for each training step are 64 and 1, respectively. The maximum memory size for MSFIRC and CI-MSFIRC is 5000. Since this paper focuses on lifelong generative modelling under TFCL, we follow from (Ye and Bors 2020a) which adopts the Fréchet Inception Distance (FID) (Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016) to evaluate the performance of various models. The detailed setting is provided in **Appendix-H** from SM[1].

Baselines : The majority of the existing lifelong learning ap-

Methods	CelebA	3D-Chair	Average	No
finetune	35.79	9.90	22.85	1
Reservoir	20.82	11.04	15.93	1
LTS	20.68	11.47	16.07	1
LGM	21.58	11.84	16.71	1
CN-DPM	20.19	11.45	15.82	11
CGKD-GAN	18.05	11.32	**14.68**	3
CGKD-VAE	20.63	11.79	16.21	5

Table 2: FID evaluation for various models under the CelebA-Chair learning setting.

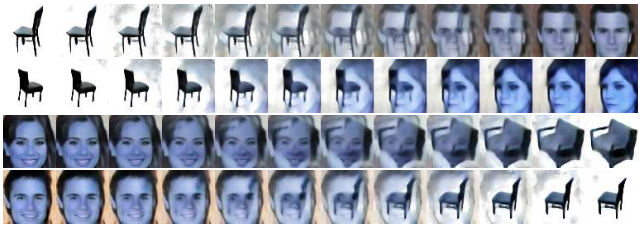

Figure 4: Image interpolation results of CGKD-GAN under CelebA-Chair setting.

proaches do not focus on generative modelling or require accessing the task information during training. Therefore, we compare the proposed approach with more related methods such as: Reservoir (Vitter 1985), Lifelong Teacher Student (LTS) (Ye and Bors 2022e), Lifelong Generative Modelling (LGM) (Ramapuram, Gregorova, and Kalousis 2017), and CN-DPM (Lee et al. 2020), respectively. For a fair comparison with CN-DPM, we also train a Student model to learn generated data by CN-DPM and from the memory buffer.

Generative Modelling Task Under TFCL

The FID results for MSFIRC are shown in Table 1, where 'No' in the last column stands for the number of experts after lifelong learning. CGKD-GAN and CGKD-VAE represent using GAN and VAE as experts in the Teacher module and '*' indicates employing the proposed Expert Pruning mechanism, according to Eq. (7). The IS result is provided in **Appendix-H2** of SM[1]. We can observe that GAN-based approaches significantly outperform VAE-based methods in terms of FID, as 67.7 by LTS versus 131.3 by LGM. The proposed CGKD-GAN outperforms CN-DPM, despite the latter using many more experts in its Teacher module. Overall, GAN-based approaches can provide high-quality generative replay patterns compared to the VAE-based models and thus achieve better lifelong learning performance. The image reconstruction results by CGKD-GAN are sharper than most of the baselines, as we can observe in Fig. 3. Also CGKD-GAN outperforms other baselines in the class incremental setting of the lifelong learning of CI-MSFIRC, according to the results from Table 3.

Learning Complex Data Stream Under TFCL

We examine the performance for the datasets consisting of complex images. Following from (Ye and Bors 2022e), we consider 5000 samples for testing from each database, CelebA (Liu et al. 2015) and 3D-chair (Aubry et al. 2014), and we create a data stream named CelebA-Chair consisting of these training samples. We adopt the setting of MSFIRC for CelebA-Chair and the results provided in Table 2 show that the proposed approach is better than the other methods.

Methods	MNIST	SVHN	Fashion	IFashion	RMNIST	CIFAR10	Average	No
finetune	158.1	167.6	246.2	233.3	138.6	229.4	195.6	1
Reservoir	141.7	163.6	220.0	200.1	127.1	115.5	161.3	1
LTS	101.9	99.4	140.6	139.5	99.9	95.5	112.8	1
LGM	108.5	122.1	189.5	175.9	96.6	92.4	130.9	1
CN-DPM	90.9	62.0	109.0	95.0	77.9	95.5	88.4	18
CGKD-GAN	16.7	65.1	44.5	43.9	27.9	85.2	**47.2**	11
CGKD-VAE	102.6	69.9	117.1	99.5	113.0	82.7	97.5	11
CGKD*-GAN	13.5	72.7	89.9	52.1	12.4	71.9	52.1	7
CGKD*-VAE	131.0	70.3	106.7	92.2	126.5	87.7	102.4	7

Table 3: FID evaluation under the CI-MSFIRC setting.

We also investigate the ability of the Student module to learn cross-domain interpolations in a single latent space. After the lifelong learning, we perform interpolations on the latent space and the visual results are shown in Fig. 4. We observe that a 3D chair can be seamlessly transformed into a human face, with the outline of the chair gradually becoming the eyes of a person. These results show that the Student can learn cross-domain latent representations under TFCL and would implicitly model the correlations between different regions of two data domains into a single latent space.

Ablation Study

We first examine the performance of the proposed CGKD-GAN when varying the threshold ν in Eq. (1). The average FID score is provided in Fig. 2, where the result shows no significant change for different ν. A small ν tends to result in adding more experts for the Teacher module. This result shows that more experts do not lead to greater performance gains, and an appropriate ν would represent a trade-off between model complexity and performance. Additional ablation studies are provided in **Appendix-I** of SM[1]. In addition, we also extend the proposed framework to the classification task (See **Appendix-I.8** from SM[1]) and explore another dynamic expansion mechanism (See **Appendix-I.10** from SM[1]).

Conclusion

In this paper, we propose a new framework for task-agnostic lifelong generative modelling from several different data domains without forgetting. We introduce the Knowledge Incremental Assimilation Mechanism (KIAM) to progressively increase the Teacher's knowledge, resulting in a model with a minimal number of parameters. To enable the Student to learn cross-domain representations, we introduce a new data-free approach that transfers the Teacher's knowledge to the Student without accessing any past samples. For maintaining a compact Teacher structure we propose a KD pruning approach for removing those experts with overlapping probabilistic representations.

References

Achille, A.; Eccles, T.; Matthey, L.; Burgess, C.; Watters, N.; Lerchner, A.; and Higgins, I. 2018. Life-long disentangled representation learning with cross-domain latent homologies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 9873–9883.

Aljundi, R.; Belilovsky, E.; Tuytelaars, T.; Charlin, L.; Caccia, M.; Lin, M.; and Page-Caccia, L. 2019a. Online Continual Learning with Maximal Interfered Retrieval. In *Advances in Neural Information Processing Systems (NeurIPS), arXiv preprint arXiv:1908.04742*.

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

Aljundi, R.; Lin, M.; Goujaud, B.; and Bengio, Y. 2019b. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems (NeurIPS), arXiv preprint arXiv:1903.08671*.

Aubry, M.; Maturana, D.; Efros, A. A.; Russell, B. C.; and Sivic, J. 2014. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3762–3769.

Banayeeanzade, M.; Mirzaiezadeh, R.; Hasani, H.; and Soleymani, M. 2021. Generative vs. Discriminative: Rethinking The Meta-Continual Learning. *Advances in Neural Information Processing Systems*, 34.

Burda, Y.; Grosse, R.; and Salakhutdinov, R. 2015. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark Experience for General Continual Learning: a Strong, Simple Baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*.

De Lange, M.; and Tuytelaars, T. 2021. Continual prototype evolution: Learning online from non-stationary data streams. In *Proc. of the IEEE/CVF International Conference on Computer Vision*, 8250–8259.

Domke, J.; and Sheldon, D. R. 2018. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, 4470–4479.

Egorov, E.; Kuzina, A.; and Burnaev, E. 2021. BooVAE: Boosting Approach for Continual Learning of VAE. *Advances in Neural Information Processing Systems*, 34.

Fernando, C.; Banarse, D.; Blundell, C.; Zwols, Y.; Ha, D.; Rusu, A. A.; Pritzel, A.; and Wierstra, D. 2017. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

Golkar, S.; Kagan, M.; and Cho, K. 2019. Continual Learning via Neural Pruning. *CoRR*, abs/1903.04476.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of Wasserstein GANs. In *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 5767–5777.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2019. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3771–3778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 6626–6637.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In *Proc. NIPS Deep Learning Workshop, arXiv preprint arXiv:1503.02531*.

Hung, C.-Y.; Tu, C.-H.; Wu, C.-E.; Chen, C.-H.; Chan, Y.-M.; and Chen, C.-S. 2019. Compacting, Picking and Growing for Unforgetting Continual Learning. In *Advances in Neural Information Processing Systems*, 13647–13657.

Jin, X.; Sadhu, A.; Du, J.; and Ren, X. 2021. Gradient-based Editing of Memory Examples for Online Task-free Continual Learning. In *Advances in Neural Information Processing Systems (NeurIPS), arXiv preprint arXiv:2006.15294*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017. Overcoming catastrophic forgetting in neural networks. *Proc. of the National Academy of Sciences (PNAS)*, 114(13): 3521–3526.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto.

Kurle, R.; Cseke, B.; Klushyn, A.; van der Smagt, P.; and Günnemann, S. 2020. Continual Learning with Bayesian Neural Networks for Non-Stationary Data. In *Proc. Int. Conf. on Learning Representations (ICLR)*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11): 2278–2324.

Lee, S.; Ha, J.; Zhang, D.; and Kim, G. 2020. A Neural Dirichlet Process Mixture Model for Task-Free Continual Learning. In *Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2001.00689*.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(12): 2935–2947.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 3730–3738.

Nam, G.; Yoon, J.; Lee, Y.; and Lee, J. 2021. Diversity Matters When Learning From Ensembles. *Advances in Neural Information Processing Systems*, 34.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Nguyen, C. V.; Li, Y.; Bui, T. D.; and Turner, R. E. 2018. Variational continual learning. In *Proc. of Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1710.10628*.

Oring, A.; Yakhini, Z.; and Hel-Or, Y. 2021. Autoencoder Image Interpolation by Shaping the Latent Space. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8281–8290. PMLR.

Phuong, M.; and Lampert, C. 2019. Towards Understanding Knowledge Distillation. In *International Conference on Machine Learning*, 5142–5151.

Polikar, R.; Upda, L.; Upda, S. S.; and Honavar, V. 2001. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE Trans. on Systems Man and Cybernetics, Part C*, 31(4): 497–508.

Ramapuram, J.; Gregorova, M.; and Kalousis, A. 2017. Lifelong generative modeling. In *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1705.09847*.

Rao, D.; Visin, F.; Rusu, A. A.; Teh, Y. W.; Pascanu, R.; and Hadsell, R. 2019. Continual Unsupervised Representation Learning. In *Proc. Neural Inf. Proc. Systems (NIPS)*, 7645–7655.

Ren, B.; Wang, H.; Li, J.; and Gao, H. 2017. Life-long learning based on dynamic combination model. *Applied Soft Computing*, 56: 398–404.

Ritter, H.; Botev, A.; and Barber, D. 2018. Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 3742–3752.

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

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training GANs. In *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 2234–2242.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. In *Advances in Neural Inf. Proc. Systems (NIPS)*, 2990–2999.

Sun, S.; Calandriello, D.; Hu, H.; Li, A.; and Titsias, M. 2022. Information-theoretic Online Memory Selection for Continual Learning. In *International Conference on Learning Representations (ICLR), arXiv preprint arXiv:2204.04763*.

Vitter, J. S. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1): 37–57.

Wen, Y.; Tran, D.; and Ba, J. 2020. BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning. In *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:2002.06715*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Ye, F.; and Bors, A. G. 2020a. Learning Latent Representations Across Multiple Data Domains Using Lifelong VAEGAN. In *Proc. European Conf. on Computer Vision (ECCV), vol. LNCS 12365*, 777–795.

Ye, F.; and Bors, A. G. 2020b. Lifelong learning of interpretable image representations. In *Proc. Int. Conf. on Image Processing Theory, Tools and Applications (IPTA)*, 1–6.

Ye, F.; and Bors, A. G. 2021a. Lifelong Infinite Mixture Model Based on Knowledge-Driven Dirichlet Process. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 10695–10704.

Ye, F.; and Bors, A. G. 2021b. Lifelong Twin Generative Adversarial Networks. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 1289–1293.

Ye, F.; and Bors, A. G. 2022a. Continual variational autoencoder learning via online cooperative memorization. In *Proc. European Conference on Computer Vision (ECCV), vol. LNCS 13683*, 531–549.

Ye, F.; and Bors, A. G. 2022b. Dynamic Self-Supervised Teacher-Student Network Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ye, F.; and Bors, A. G. 2022c. Learning an evolved mixture model for task-free continual learning. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 1936–1940.

Ye, F.; and Bors, A. G. 2022d. Lifelong Generative Modelling Using Dynamic Expansion Graph Model. In *Proc. AAAI on Artificial Intelligence*, 8857–8865.

Ye, F.; and Bors, A. G. 2022e. Lifelong Teacher-Student Network Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6280–6296.

Ye, F.; and Bors, A. G. 2022f. Task-Free Continual Learning via Online Discrepancy Distance Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ye, F.; and Bors, A. G. 2023. Lifelong Mixture of Variational Autoencoders. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1): 461–474.

Yoon, J.; Madaan, D.; Yang, E.; and Hwang, S. J. 2022. Online Coreset Selection for Rehearsal-based Continual Learning. In *International Conference on Learning Representations (ICLR), arXiv preprint arXiv:2106.01085*.

Zhai, M.; Chen, L.; Tung, F.; He, J.; Nawhal, M.; and Mori, G. 2019. Lifelong GAN: Continual Learning for Conditional Image Generation. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2759–2768.