Lifelong learning of interpretable image representations

Fei Ye and Adrian G. Bors

Department of Computer Science, University of York, York YO10 5GH, UK

***Abstract*—Existing machine learning systems are trained to adapt to a single database and their ability to acquire additional information is limited. Catastrophic forgetting occurs in all deep learning systems when attempting to train them with additional databases. The information learnt previously is forgotten and no longer recognized when such a learning systems is trained using a new database. In this paper, we develop a new image generation approach defined under the lifelong learning framework which prevents forgetting. We employ the mutual information maximization between the latent variable space and the outputs of the generator network in order to learn interpretable representations, when learning using the data from a series of databases sequentially. We also provide the theoretical framework for the generative replay mechanism, under the lifelong learning setting. We perform a series of experiments showing that the proposed approach is able to learn a set of disjoint data distributions in a sequential manner while also capturing meaningful data representations across domains.**

***Index Terms*—Lifelong learning, Representation learning, Generative Adversarial Networks, Mutual information.**

I. Introduction

One inherent advantage of humans and animals is that of being able to continually acquire new skills, by learning progressively while aging, without forgetting the previously learnt knowledge throughout their lifespan [1]. However, artificial learning systems when learning from the data sampled from successive databases, have their parameters tuned onto the probabilistic representation of the latest available database, while forgetting the previously learnt information. The reason for this outcome, present in all existing systems requiring training, is that their objective function is designed to tune the network parameters by optimizing a match between target labels and network outputs. Meanwhile, after training for a new task using backpropagation, the new model would forget completely the previously learnt knowledge.

Many of the previous approaches, aiming to address catastrophic forgetting, often focus on implementing dynamic network systems [2], [3]. Such systems would aim to increase the number of layers and processing units on each layer in order to acquire additional information. Other attempts to address this issue would impose a large penalty in the objective function that prevents from significantly changing the network parameters while learning a new task [4]–[7]. However, such approaches are often sensitive to the choice of data being learnt and lead to significant increases in the required computational resources. Hanul *et al.* [8] introduces a dual architecture consisting of a powerful generator and a classifier. Adversarial learning is used to train the generator with the accumulated data produced by the generator after learning the previously given tasks. Similar works based on the Generative Adversarial Networks (GAN) [9] framework are also proposed in [10]–[12]. The inference abilities of Variational Autoencoders (VAE) have been combined with the generation capability of GANs in the lifelong VAEGAN learning [13].

This paper has the following contributions: (1) We propose a novel GAN based framework for lifelong learning. (2) We introduce a theoretical probabilistic framework for the generative replay mechanism used in the context of lifelong learning. (3) We employ the mutual information maximization between the latent variables and generator outputs under the lifelong learning framework, in order to enable a mechanism for capturing meaningful data representations across domains. The methodology and theoretical framework for the proposed lifelong interpretable learning framework is described in Section II. Experimental results are provided in Section III and the conclusions of this study are drawn in Section IV.

II. Lifelong interpretable learning

In this section, we introduce a novel lifelong framework which besides aiming to generate high quality images also captures interpretable representations across data domains. We consider that each data domain characterizes a distinct database. Let us consider a set of databases, each characterized by a data distribution, $\mathbf{o}_1 \sim p(\mathcal{O}_1), \mathbf{o}_2 \sim p(\mathcal{O}_2), \ldots, \mathbf{o}_K \sim p(\mathcal{O}_K)$, which are being learnt in a sequential manner. We consider that $\mathbf{o}_i$ is an image sampled from the target database defined by $p(\mathcal{O}_i)$, $i = 1, \ldots, K$. Unlike in the traditional lifelong learning tasks which aim to make predictions from all learnt data samples, in this study we seek to define meaningful data representations that can interpret data characteristics. In the lifelong learning problem, each data distribution $p(\mathcal{O})$, where $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_K\}$ is only seen once. The goal of the proposed algorithm is to learn a model, which is able to generate all images from the previously seen databases, while capturing meaningful image representations and characteristics across domains.

978-1-7281-8750-1/20/$31.00 ©2020 IEEE

A. Training the data generator for a single task

After learning a single data distribution, we assess its generative replay capability. Such generative mechanisms can be used for learning successively a series of distributions under the lifelong setting, without the need to see each time the real data. Let us consider that $\mathbf{o}$ represents the observed data sampled from the target distribution $p(\mathcal{O})$, and $\{\mathbf{z}, \mathbf{c}, \mathbf{d}\}$ are three independent random vectors, representing random noise, continuous latent variables and discrete variables, sampled from the prior distributions $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{c} \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{d} \sim Cat(k, 1/k)$, where the first two distributions are Gaussian while the third denotes a categorical distribution which is specific for discrete variables. We want to train a generator network $G(\mathbf{z}, \mathbf{c}, \mathbf{d})$ to approximate the data distribution $p(\mathcal{O})$ by using adversarial learning. Adversarial learning is defined by Min-Max optimization:

$$\min_G \max_{D \in \Phi} \left[\mathbb{E}_{\mathbf{o} \sim p(\mathcal{O})}[D(\mathbf{o})] - \mathbb{E}_{\mathbf{o}' \sim p_G(\mathcal{O}')}[D(\mathbf{o}')]\right] \quad (1)$$

where we use the Earth mover distance optimisation as in the Wasserstein GAN model [14], [15], instead of the Jensen-Shannon divergence [9], for measuring the distance between the true data $p(\mathcal{O})$ and the probability of the generated data $\mathbf{o}' \sim p_G(\mathcal{O}')$. Φ denotes a set of 1-Lipschitz functions. We further consider using the gradient penalty proposed in [16] in order to enforce the Lipschitz constraint, resulting in:

$$\begin{aligned} &\min_G \max_{D \in \Phi} \left[\mathbb{E}_{\mathbf{o} \sim p(\mathcal{O})}[D(\mathbf{o})] - \mathbb{E}_{\mathbf{o}' \sim p_G(\mathcal{O})}[D(\mathbf{o}')]\right] \\ &+ \lambda \mathbb{E}_{\tilde{o} \sim \mathbb{P}_{\tilde{o}}} \left[\left(\|\nabla_{\tilde{o}} D\left(\tilde{o}\right)\|_2 - 1\right)^2\right]. \end{aligned} \quad (2)$$

While this objective function is used for learning a single task, in the following we show how generative replay mechanisms can be used for training the model with multiple tasks, each learned from data sampled from a different database, $\mathbf{o}_i \sim p(\mathcal{O}_i)$, $i = 1, \ldots, k$.

B. The theoretic framework for the generative replay mechanism in the context of lifelong learning

In this section, we provide the theoretic analysis for the generative replay mechanism used for the lifelong learning in artificial systems.

Definition 1. Let us consider that $p(\mathcal{O}) = \prod_{i=1}^{k} p(\mathcal{O}_i)$ represents the true joint data distribution in which each individual dataset $p(\mathcal{O}_i)$ is assumed to be independent from the others.

Definition 2. Let us consider that $p(\hat{\mathcal{O}}_k)$ represents the output fake data distribution produced by the generator network $G_{\omega_k}(\mathbf{c}, \mathbf{d}, \mathbf{z})$, after training with the data corresponding to the k-th task, where ω_k represents the network's parameters. During the k-th task learning, we consider that only $p(\mathcal{O}_{k-1})$ and $p(\hat{\mathcal{O}}_{k-1})$ are available under the lifelong training setting :

$$p(\hat{\mathcal{O}}_k | \hat{\mathcal{O}}_{k-1}, \mathcal{O}_k) = 1 - \min(\|W(p(\hat{\mathcal{O}}_k), p(\hat{\mathcal{O}}_{k-1}, \mathcal{O}_k))\|, 1) \quad (3)$$

where $W(\cdot)$ is the Wasserstein distance, representing the Earth mover distance optimization [14], [15]. The expression from (3) represents the probability of $\hat{\mathcal{O}}_k$ when observing simultaneously $\hat{\mathcal{O}}_{k-1}$ and $\hat{\mathcal{O}}_k$, and :

$$p(\hat{\mathcal{O}}_k | \hat{\mathcal{O}}_{k-1}, \mathcal{O}_k) = 1 \rightarrow W(p(\hat{\mathcal{O}}_k), p(\hat{\mathcal{O}}_{k-1}, \mathcal{O}_k)) = 0. \quad (4)$$

Theorem 1. The information characterizing $p(\hat{\mathcal{O}}_k)$ depends on all previously learned distributions.

Proof. From the fact that $p(\hat{\mathcal{O}}_{k-1})$ is independent from $p(\mathcal{O}_k)$, we derive for the marginal probability $p(\hat{\mathcal{O}}_k)$ through mathematical induction:

$$\begin{aligned} p(\hat{\mathcal{O}}_k) &= \int \int p(\hat{O}_k | \hat{\mathcal{O}}_{k-1}, \mathcal{O}_k) p(\hat{\mathcal{O}}_{k-1}) p(\mathcal{O}_k) d\hat{\mathcal{O}}_{k-1} d\mathcal{O}_k \\ &= \int \ldots \int p(\hat{\mathcal{O}}_1) \prod_{i=0}^{k-2} p(\hat{O}_{k-i} | \hat{O}_{k-1-i}, \mathcal{O}_{k-i}) \\ &\prod_{i=0}^{k-2} p(\mathcal{O}_{k-i}) d\hat{\mathcal{O}}_{k-1} \ldots d\hat{\mathcal{O}}_1 d\mathcal{O}_k \ldots d\mathcal{O}_2 \end{aligned} \quad (5)$$

where this equation is integrated over all the previously learnt data samples from all databases $\mathcal{O}_i$ $i = 1, \ldots, k$.

Lemma 1. If the learnt distribution $p(\hat{\mathcal{O}}_i)$ is an exact approximation to the target distribution when learning every task, then the latest learnt distribution $p(\hat{\mathcal{O}}_i)$ is an exact probabilistic approximation to the true joint distribution of data $\prod_{i=1}^{k} p(\mathcal{O}_i)$.

Proof. If we consider that all previously learnt distributions are exact representations of their target distributions, we have $\prod_{i=1}^{k} p(\hat{\mathcal{O}}_{k-i} | \hat{\mathcal{O}}_{k-i-1}, \mathcal{O}_{k-i}) = 1$ and $W(p(\hat{\mathcal{O}}_1), p(\mathcal{O}_1)) = 0 \rightarrow p(\hat{\mathcal{O}}_1) = p(\mathcal{O}_1)$. Then the conditional probability can be rewritten as :

$$p(\hat{\mathcal{O}}_k | \bigcup_{i=1}^{k-1} \hat{\mathcal{O}}_i, \mathcal{O}_k) = 1 \rightarrow W(p(\hat{\mathcal{O}}_k), \prod_{i=1}^{k} p(\mathcal{O}_i)) = 0, \quad (6)$$

and the Min-Max optimization becomes :

$$\begin{aligned} \min_G \max_{D \in \Phi} &\left[\mathbb{E}_{\mathbf{o} \sim p(\hat{\mathcal{O}}_{k-1}, \mathcal{O}_k)}[D(\mathbf{o})] - \mathbb{E}_{\mathbf{o}' \sim \mathbb{P}_G}[D(\mathbf{o}')]\right] \\ &+ \lambda \mathbb{E}_{\hat{\mathbf{o}} \sim \mathbb{P}_{\hat{o}}}[(\|\nabla_{\hat{\mathbf{o}}} D(\hat{o})\|_2 - 1)^2]. \end{aligned} \quad (7)$$

Lemma 2. The necessary and sufficient condition to have a good representation for all databases is to approximate well each database during the lifelong learning.

Proof. By considering *argumentum ad absurdum* rhetoric, if $\prod_{i=1}^{k-2} p(\hat{\mathcal{O}}_{k-i} | \hat{\mathcal{O}}_{k-i-1}, \mathcal{O}_{k-i}) \neq 1$, then $p(\hat{\mathcal{O}}_k)$ may not be a good approximation to $\prod_{i=1}^{k} p(\mathcal{O}_i)$. However, from Definition 2, we have $p(\hat{\mathcal{O}}_k | \hat{\mathcal{O}}_{k-1}, \mathcal{O}_k) = 1$. After learning the corresponding probabilistic representation $p(\hat{\mathcal{O}}_i)$, this relies on the previously learnt distributions $p(\hat{\mathcal{O}}_{i-1})$ while also learning the new true distribution $p(\mathcal{O}_i)$, we have :

$$\prod_{i=0}^{k-2} p(\hat{\mathcal{O}}_{k-i} | \hat{\mathcal{O}}_{k-i-1}, \mathcal{O}_{k-i}) = 1 \text{ and } p(\hat{\mathcal{O}}_1) = p(\mathcal{O}_1) \quad (8)$$

in order to approximate $\prod_{i=1}^{k} p(\mathcal{O}_i)$ exactly. This contradicts the initial assumption stated above. Equation (8) indicates that given $\prod_{i=0}^{k-2} p(\hat{\mathcal{O}}_{k-i} | \hat{\mathcal{O}}_{k-i-1}, \mathcal{O}_{k-i}) = 1$ we have through $p(\hat{\mathcal{O}}_k)$ a probabilistic representation of all given databases.

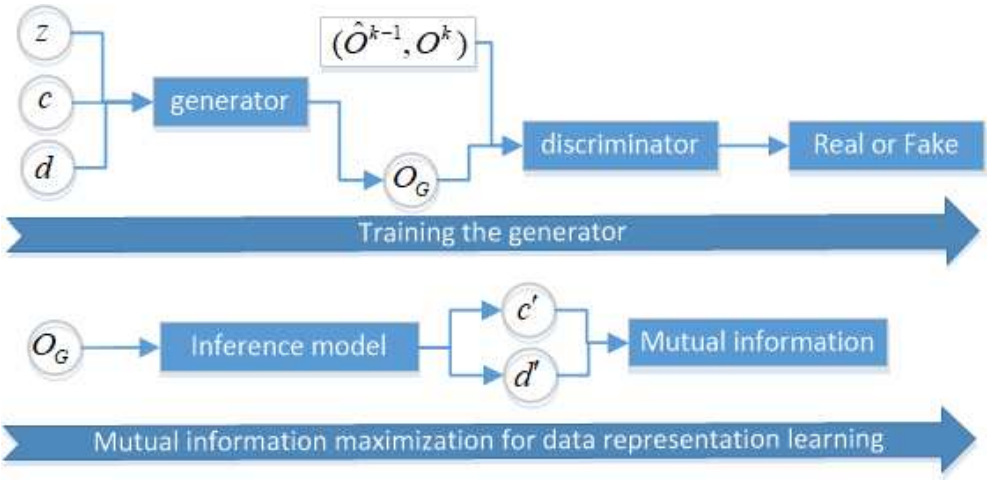


Fig. 1. The structure of the proposed lilfelong learning through the mutual information maximization model.

C. Learning data representations by mutual information maximization

In information theory, mutual information (MI) measures the amount of information shared by one random variable when observing another variable. In the proposed approach, we want to learn simultaneously discrete and continuous interpretable representations.

Let us consider firstly learning the data representation for one task from a given database. Let $\mathbf{u} = (\mathbf{d}, \mathbf{c})$ represent the joint latent variable for discrete and continuous latent variables. The learning goal of the proposed approach is to maximize the mutual information between the joint latent variables $\mathbf{u}$ and the distribution generated by $G(\mathbf{c}, \mathbf{u})$. The mutual information is defined by :

$$\mathrm{I}(\mathbf{u}, \mathrm{G}(\mathbf{z}, \mathbf{u})) = \mathrm{H}(\mathbf{u}) - \mathrm{H}(\mathbf{u}|\mathrm{G}(\mathbf{z}, \mathbf{u})) \tag{9}$$

where $H(\mathbf{u}|G(\mathbf{z}, \mathbf{u}))$ is the conditional entropy, which measures the uncertainty of estimating $\mathbf{u}$ when observing $G(\mathbf{z}, \mathbf{u})$ and $H(\mathbf{u})$ represents the entropy of the latent variables. By maximizing $\mathrm{I}(\mathbf{u}, \mathrm{G}(\mathbf{z}, \mathbf{u}))$ we can reduce this uncertainty and therefore preserve the latent information during the generation process. Similar MI objectives have been adopted in the research studies from [17]–[20]. However, it is challenging to optimize the mutual information directly, given that it depends on inferring the true posterior $p(\mathbf{u}|\mathbf{o})$. In order to address this challenge, we define an auxiliary distribution $S(\mathbf{u}|\mathbf{o})$ to approximate the true posterior $p(\mathbf{u}|\mathbf{o})$, and then we derive a lower bound on the mutual information, called $\mathcal{L}_{MI}$:

$$\begin{aligned}
&\mathrm{I}(\mathbf{u}, \mathrm{G}(\mathbf{z}, \mathbf{u})) = \int\int \mathrm{G}(\mathbf{z}, \mathbf{u}) \mathrm{p}(\mathbf{u}|\mathbf{o}) \log \frac{\mathrm{p}(\mathbf{u}|\mathbf{o})}{\mathrm{S}(\mathbf{u}|\mathbf{o})} \mathrm{d}\mathbf{o}\mathrm{d}\mathbf{u} + \\
&\int\int G(\mathbf{z}, \mathbf{u}) p(\mathbf{u}|\mathbf{o}) \log S(\mathbf{u}|\mathbf{o}) d\mathbf{o} d\mathbf{u} + H(\mathbf{u}) = \\
&= \mathbb{E}_{\mathbf{o}\sim G(\mathbf{z},\mathbf{u})}[D_{KL}[p(\mathbf{u}|\mathbf{o})||S(\mathbf{u}|\mathbf{o})]] + \\
&\mathbb{E}_{\mathbf{o}\sim G(\mathbf{z},\mathbf{u})}[\mathbb{E}_{\mathbf{u}\sim p(\mathbf{u},\mathbf{o})}[\log S(\mathbf{u}|\mathbf{o})]] + H(\mathbf{u}) \geqslant \\
&\geqslant \mathbb{E}_{\mathbf{o}\sim G(\mathbf{z},\mathbf{u})}[\mathbb{E}_{\mathbf{u}\sim p(\mathbf{u},\mathbf{o})}[\log S(\mathbf{u}|\mathbf{o})]] + H(\mathbf{u}) = \mathcal{L}_{MI}
\end{aligned} \tag{10}$$

where we consider that the KL divergence is positive or at least equal to 0. In this study, we treat $H(\mathbf{u})$ as a constant for simplicity. The auxiliary distribution $S(\mathbf{u}|\mathbf{o})$ is implemented by using a neural network with two output layers, one for producing continuous variables and the other for calculating discrete variables.

Theorem 2. The inference model $S(\mathbf{u}|\mathbf{o})$ can be used in the context of the generative reply mechanism for learning representations from multiple domains.

Lemma 3. If the generator approximates exactly its target distribution after learning every task, the inference model $S(\mathbf{u}|\mathbf{o})$ can learn from the data associated with the corresponding modes from all previously learnt probabilistic representations of the given data distributions.

Proof. From Lemma 1, we know that $p(\hat{\mathcal{O}}_k) = \prod_{i=1}^{k} p(\mathcal{O}_i)$. Considering from Definition 2 that $p(\hat{\mathcal{O}}_k)$ characterizes the distribution of the output of $G_{\omega_k}(\mathbf{c}, \mathbf{d}, \mathbf{z})$, where ω_k represents the network parameters, the inference model actually learns the probabilistic data representations from $\prod_{i=1}^{k} p(\mathcal{O}_i)$ by using the mutual information maximization, during the lifelong learning process.

The diagram for the proposed lifelong learning through mutual information maximization model is presented in Figure 1.

III. Experimental results

In the following we provide the experiments showing how the proposed model can learn interpretable representations across the domains of several databases, under the lifelong learning setting. We implement the generator, discriminator, and inference by using deep convolution neural networks (CNN). We use Tensorflow and learning through stochastic optimization using Adam [21] with a learning rate of 0.0001 for all models.

A. The lifelong learning from MNIST to MNIST-Fashion

In this section, we evaluate the performance of the proposed approach when firstly learning MNIST database [22], containing images of handwritten digits, and then MNIST-Fashion

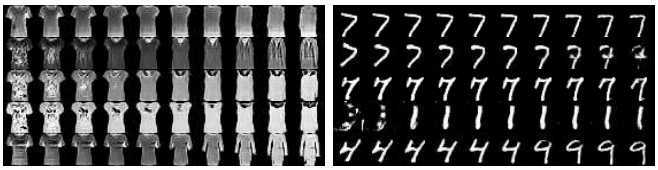
(a) Proposed lifelong learning approach.

(b) Conventional training without any generative reply.

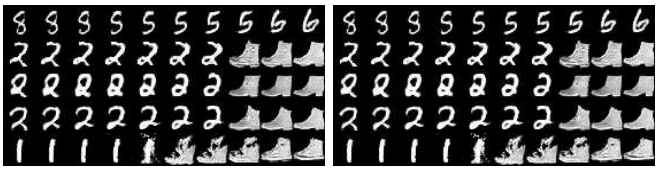
Fig. 2. Generation results when firstly learning MNIST database followed by the MNIST-Fashion.

database [23], which contains images of clothing items. These two databases, contain only greylevel images and have the same number of classes and data samples, while displaying completely different information. The generated images following the lifelong MNIST to MNIST-Fashion learning are presented in Figure 2a, while the images generated without the generative reply are shown in Figure 2b. We observe that the proposed lifelong learning approach can generate images characteristic to both data domains, unlike in the classical approach, where we do not have a generative reply mechanism.

Fig. 3. Generated results when varying the discrete variable $\mathbf{d}$ along columns and the continuous variables $\mathbf{c}$ along rows.



(a) $\mathbf{c}_1$ in MNIST-Fashion (b) $\mathbf{c}_1$ in MNIST



(c) $\mathbf{c}_2$ in MNIST (d) $\mathbf{c}_2$ in MNIST

Fig. 4. Exploring the latent space for MNIST and Fashion and databases, under the MNIST to Fashion lifelong learning, where we change a single latent variables from -1.0 to 1.0 while fixing the others.

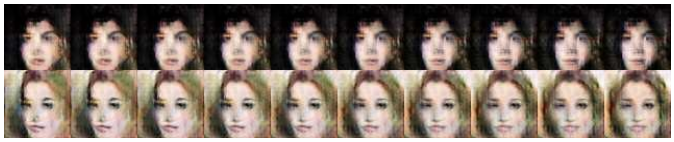
(a) The proposed lifelong approach.

(b) Classical approach.

Fig. 5. Generation results after the lifelong learning from CelebA to 3D-Chairs databases.

(a) Gender

(b) Makeup

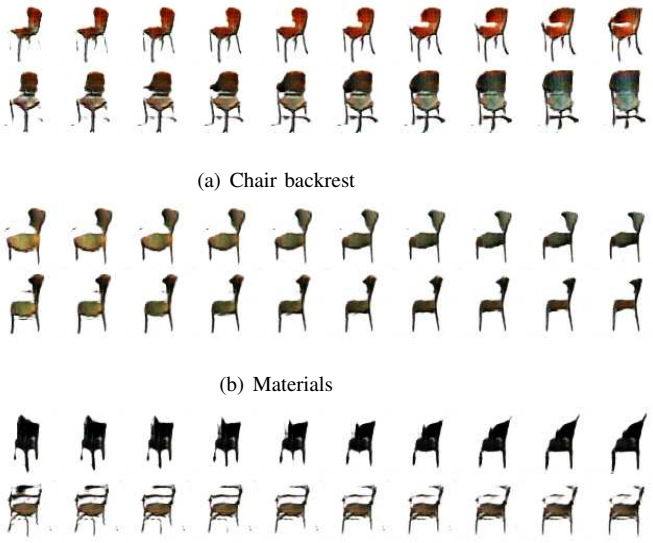


(c) Face orientation

Fig. 6. Results when manipulating characteristics in images of faces from the CelebA dataset, under the CelebA to 3D-chairs lifelong learning.

B. Disentangled representations under the lifelong learning using the mutual information maximization

In the following we examine the disentanglement ability of the proposed lifelong learning using the mutual information maximization within the latent space of the generated images. We fix the continuous latent variables and change the discrete variable from 0 to 9. The generated images are shown in Figure 3, where each column is produced considering the same discrete variable while the images from each row correspond to a different continuous variable. We observe that this model is able to capture the discriminating attributes from both MNIST and MNIST-Fashion databases without any mixing between the data from the two databases. We then fix other variables and change two continuous latent variables, $\mathbf{c}_1$ and $\mathbf{c}_2$ from -1 to 1 and the results are shown in Figures 4a-d. From these results we can observe that the proposed approach can capture independently various clothing styles from MNIST-Fashion and the writing styles of the digits from MNIST. Meanwhile, when changing one of the continuous variable, the resulting images would interpolate between a digit and a shoe, as it can



(a) Chair backrest

(b) Materials

(c) Orientation by changing the azimuth

Fig. 7. Results when manipulating characteristics in the 3D-chairs images under the CelebA to 3D-chairs lifelong.

be observed from Figures 4c and 4d.

C. The lifelong learning from CelebA to 3D-chairs databases

In this section, we evaluate the performance of the proposed approach on CelebA [24], and 3D-chairs [25] databases, containing face images of well known persons (celebrities) and 3-D chairs, respectively. We train the proposed model under the CelebA to 3D-chairs lifelong learning based on the mutual information maximization. The results are presented in Figure 5a, where it is shown that the proposed approach can generate images from the domains of both databases. In Figure 5b, we show the results when considering the classical approach without using the generative replay mechanism. From these images it can be observed that this model forgets quickly the images from the previously learned tasks, such as the images of faces characteristic to CelebA database.

In another experiment we modify a single continuous latent variable between -1 and 1, while fixing all others during the generation process. The results for the face images from CelebA, shown in Figures 6a, 6b and 6c, demonstrate that the proposed lifelong learning approach is able to discover disentangled representations characteristic of gender, makeup change and face orientation change, respectively. Meanwhile, the results for the 3D-chairs are shown in Figures 7a, 7b and 7c where we show how the proposed model learns disentangled representations characteristic of chair backrest, material type and for changing the azimuth in the chairs' orientation, respectively.

D. Numerical evaluations

In this section, we use the Fréchet Inception Distance (FID) [26] in order to evaluate the quality of the generated image results under the lifelong learning. We train the proposed lifelong learning model under the CelebA to 3D-Chair learning. FID is evaluated on both CelebA and 3D-Chair images, and the numerical results are provided in the bar-plot from Figure 8, where we compare the proposed model with other lifelong learning approaches, such as LGAN [27] and LGM [28]. These results indicate that the proposed model generates images of similar quality when compared to those generated by LGAN. Moreover, unlike LGAN, the proposed model is able to learn disentangled representation across domains under the lifelong learning. Meanwhile, the proposed model is able to produce higher-quality generative replay images than LGM.

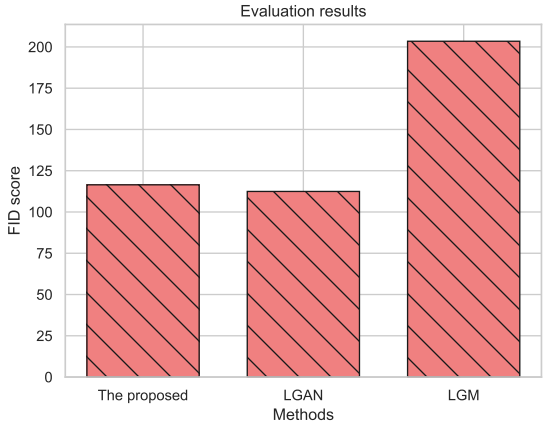


Fig. 8. FID results for the generated images after the CelebA to 3D-chair lifelong learning.

IV. Conclusions

We propose a new approach for lifelong learning interpretable representations across data domains using the mutual information maximization criterion. In this approach we employ the generative replay mechanism in order to prevent forgetting the previously learnt knowledge. In order to learn interpretable representations, we maximize the mutual information between the latent representation and the generator's outputs. The theoretical analysis shows that by using a powerful generator for the data replay, the inference model can learn data representations across multiple domains. The experiments performed achieve data interpolations across different data domains. In further research work we are considering expanding the lifelong learning model to learning multiple databases while also improving the quality of the generated images.

Acknowledgement

The authors would like to thank NVIDIA for granting a Titan XP GPU, which was used for the experiments.

REFERENCES

[1] G. I. Parisi, J. Kemker, R.and Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[2] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *Proc. Int. Conf. on Learning Representations (ICLR),arXiv preprint arXiv:1708.01547*, 2017.

[3] C. Tessler, S. Givony, T. Zahavy, D. Mankowitz, and S. Mannor, "A deep hierarchical approach to lifelong learning in minecraft," in *Proc. AAAI Conf. on Artif. Intel.*, 2017, pp. 1553–1561.

[4] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[5] H. Jung, J. Ju, M. Jung, and J. Kim, "Less-forgetting learning in deep neural networks," *arXiv preprint arXiv:1607.00122*, 2016.

[6] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proc. of the Nat. Academy of Sciences (PNAS)*, vol. 114, no. 13, pp. 3521–3526, 2017.

[7] R. Polikar, L. Upda, S. Upda, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Trans. on Systems Man and Cybernetics, Part C*, vol. 31, no. 4, pp. 497–508, 2001.

[8] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Proc. Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2017, pp. 2990–2999.

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

[10] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, and G. Mori, "Lifelong GAN: Continual learning for conditional image generation," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 2759–2768.

[11] C. Wu, L. Herranz, X. Liu, J. van de Weijer, and B. Raducanu, "Memory replay GANs: Learning to generate new categories without forgetting," in *Proc. Advances In Neural Inf. Proc. Systems (NeurIPS)*, 2018, pp. 5962–5972.

[12] A. Seff, A. Beatson, D. Suo, and H. Liu, "Continual learning in generative adversarial nets," *arXiv preprint arXiv:1705.08395*, 2017.

[13] F. Ye and A. G. Bors, "Learning latent representations across multiple data domains using lifelong VAEGAN," in *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[14] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. on Machine Learning (ICML), vol. PMLR 70*, 2017, pp. 214–223.

[15] ——, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2017, pp. 5767–5777.

[17] D. Barber and F. Agakov, "The IM algorithm: a variational approach to information maximization," *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, pp. 201–208, 2003.

[18] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Advances in Neural Inf. Proc. Systems (NeurIPS)*, 2016, pp. 2172–2180.

[19] A. Krause, P. Perona, and R. Gomes, "Discriminative clustering by regularized information maximization," in *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 2010, pp. 775–783.

[20] J. Bridle, A. Heading, and D. MacKay, "Unsupervised classifiers, mutual information and 'phantom targets," in *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, 1992, pp. 1096–1101.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representations (ICLR), arXiv preprint arXiv:1412.6980*, 2015.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

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

[25] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models," in *Proc. of IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR)*, 2014, pp. 3762–3769.

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

[27] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. of Int. Conf. on Machine Learning (ICML), vol. PMLR 70*, 2017, pp. 3987–3995.

[28] J. Ramapuram, M. Gregorova, and A. Kalousis, "Lifelong generative modeling," *arXiv preprint arXiv:1705.09847*, 2017.