Study and Analysis of GAN and VAE in continual learning

Ayush Pande
20111404

Narein Rao
20111414

Pragya Agrawal
20211402

Sri Madhan M
20111420

1 Introduction

Lifelong learners are known to retain and reuse learned behavior acquired over past tasks and aim to maximize performance across all tasks. Neural networks face the problem of completely forgetting previously learned information upon learning new information, aka catastrophic forgetting. This problem of catastrophic forgetting is a central obstacle that needs to be resolved to build an effective neural lifelong learner. Existing lifelong learning methods have been roughly grouped into three categories, i.e., regularization-based, dynamic-model-based, and generative-replay-based methods. Generative replay is believed to be an effective and general strategy for lifelong learning problems, although most existing methods of this type often have blurry/distorted generation (for images) or scalability issues. The use of Generative Adversarial Networks and Variational Auto-encoders have shown to produce commendable results, maintaining the principle of continual learning and generating good images. Although it has been experimentally confirmed that GANs can produce higher quality images than VAEs, we would like to study and discover the pros and cons of each approach which may allow us to further integrate the best of both approaches.
To do so, we would perform the following analysis,

- Reproduce experimental results and gain a measure of how realistic the generated images are across the two approaches.
- Measure the forgetfulness of the two approaches, find the relationship between number of tasks and forgetfulness and compare the performance of the two approaches in continual learning.
- Generate images over a domain perceptually-distant from the domain the models have been trained over with varying sample sizes (few-shot learning)

2 Literature Review

2.1 Cong, Yulai et al. "GAN Memory with No Forgetting."

A generative adversarial network designed by Ian Goodfellow [2], wherein two neural networks (generator and discriminator) contest with each other in the form of a zero-sum game (ones loss is anothers gain). The crux of the GAN is the training that occurs dynamically, i.e. both generator and discriminator train whilst adversially outperform each other.

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim q_{data}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log(1 - D(G(\mathbf{z})))\right]$$

Furthermore, various literature were considered in the paper on the manipulation of the style

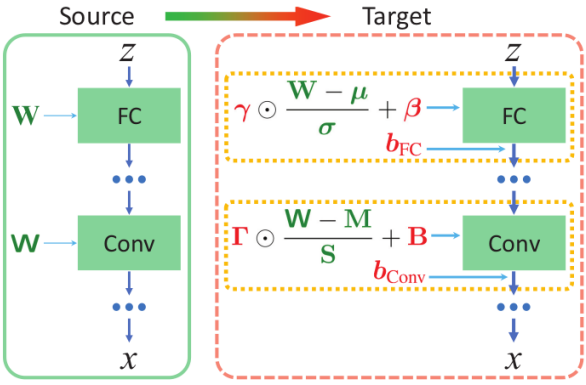


Figure 1: Depiction of the GAN memory (Src: [1])

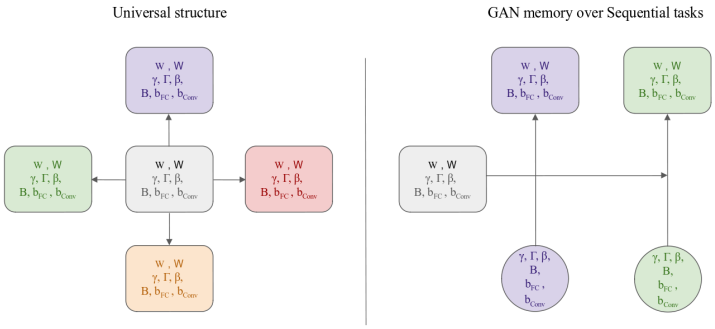


Figure 2: Left: A common universal structure amongst different images. A source GAN can be modulated to generate images over perceptually-distant target domains. Right: GAN memory over a stream of tasks.

of an image by modulating the statistics of its latent features. The paper focuses on Feature-wise Linear Modulation (FiLM), which imposes an element-wise affine transformation to the latent features of a neural network. Given a d-dimensional feature h $\in \mathbb{R}^d$ from a layer of a neural network, FiLM yields

$$\hat{h} = \gamma \odot h + \beta$$

where $\hat{h}$ is forwarded to the next layer, denotes the Hadamard product, and $\gamma \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ are the scale and shift respectively. Another technique named adaptive filter modulation (AdaFM) was also considered to modulate the source convolutional filters to and deliver a boosted performance. Given a convolution filter $\mathbf{W} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times K_1 \times K_2}$, where $C_{\text{in}}/C_{\text{out}}$ denotes the number of input/output channels and $K_1 \times K_2$ is the kernel size, AdaFM yields

$$\hat{\mathbf{W}} = \mathbf{\Gamma} \odot \mathbf{W} + \mathbf{B}$$

where the scale matrix $\mathbf{\Gamma} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ and $\mathbf{B} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ is the shift matrix.

The paper exploits GANs to produce a "GAN memory", which utilizes a modified FiLM and AdaFM (mFiLM and mAdaFM) to modulate the source to a particular target image. The mFiLM

and mAdaFM are as follows,

$$\hat{\mathbf{W}} = \boldsymbol{\gamma} \odot \frac{\mathbf{W}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} + \boldsymbol{\beta}, \quad \hat{\boldsymbol{b}} = \boldsymbol{b} + \boldsymbol{b}_{\text{FC}}$$

where $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^{d_{\text{out}}}$, with the elements $\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i$ denoting the mean and standard derivation of the vector $\mathbf{W}_{i,:}$,respectively; $\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{b}_{\text{FC}} \in \mathbb{R}^{d_{\text{out}}}$ are target-specific scale, shift, and bias style parameters

$$\hat{\mathbf{W}} = \boldsymbol{\Gamma} \odot \frac{\mathbf{W}-\mathbf{M}}{\mathbf{S}} + \mathbf{B}, \quad \hat{\boldsymbol{b}} = \boldsymbol{b} + \boldsymbol{b}_{\text{Conv}}$$

where $\mathbf{M}$,$\mathbf{S} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ with the elements $\mathbf{M}_{i,j}, \mathbf{S}_{i,j}$ denoting the mean and standard derivation of the vector vec$(\mathbf{W}_{i,j,:,:})$, respectively.The trainable target-specific style parameters are $\boldsymbol{\Gamma}, \mathbf{B} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ and $\boldsymbol{b}_{\text{Conv}} \in \mathbb{R}^{C_{\text{out}}}$

The GAN memory intuitively uses the style parameters to modify textural/structural information, control color information and control the illumination and localized objects respectively within the image. The modulation first starts by introducing changes in the overall contrast and illumination of the generation, followed by visible changes in the style as the style parameters are updated continuously and finally refining the generation details. The normalization (as seen in the above equations) helps provide better training efficiency and performance. From the above GAN memory formulation, we can see that it is possible to modulate a source GAN to any desired and perceptually distant target, which shows to hold a universal structure for images wherein it is possible to modulate a well trained GAN to generate images for any target domain provided one relatively large dataset (i.e. a common universal structure for images).

2.2 Ramapuram, Jason et al. "Lifelong Generative Modeling."

Standard Auto-Encoder uses two deep neural networks - encoder and decoder network. Encoder network produces new feature representation (called as latent space) from the given input, whereas the decoder tries to reproduces the original given input from the latent space representation. Variational Auto-Encoder is similar to Auto-Encoder, but instead of simply producing the latent space and reconstructing input, the input will be encoded as the distribution over the latent space representation and then a point is sampled from this distribution which is used to reconstruct the original input.

The parameters of the latent variable models are typically optimized by maximizing the marginal likelihood estimate $\max_\theta P_\theta(x) = \max_\theta \int P_\theta(x|z)P(z)dz$, where $P_\theta(x|z)$ is the likelihood. Since the posterior $P_\theta(z|x)$ is generally intractable, variational inference is used to approximate the intractable posterior using a known distribution $Q_\phi(z|x)$

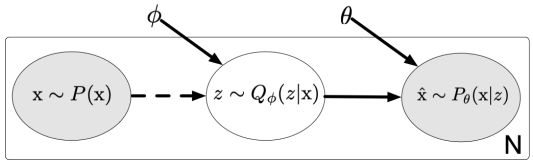


Figure 3: Standard VAE (src:[4])

$$\begin{aligned}\log P_\theta(x) &= \log \int \frac{Q_\phi(z \mid x)}{Q_\phi(z \mid x)} P_\theta(x \mid z) P(z) dz \\ &\geq \mathbb{E}_Q\big[\log P_\theta(x \mid z)\big] - \mathbb{D}_{KL}\big[Q_\phi(z|x) \mid\mid P(z)\big]\end{aligned}$$

The above inequality equation is called as evidence lower bound (ELBO), and the parameters of the VAE are optimized by maximizing the ELBO function.

For continual learning, student teacher architecture is used in [4]. The role of the teacher model is to act as probabilistic knowledge store of learned distribution of previously trained dataset, and the student model learns this knowledge store from teacher in the form of generative replay.

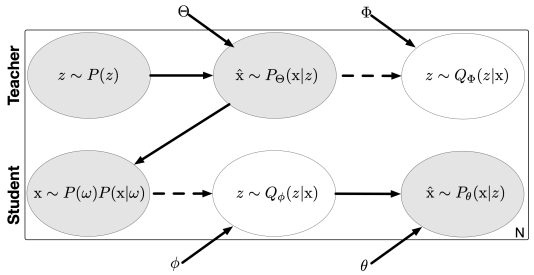


Figure 4: Student-teacher architecture for Life-long VAE (src:[4])

In the student teacher architecture when a new dataset comes, the current student model is frozen and promoted as new teacher, while the old teacher model is dropped. A new student model is then created with latest weights (ϕ and θ) of the previously used student model. And for learning this new dataset, student receives the data from current dataset and synthetically generated data from the newly promoted teacher model. Because of this cyclic process, there is only 2 VAE models at any given time. Here, ϕ and θ are the parameters of the student model's encoder and decoder, respectively. Similarly, Φ and Θ are the parameters of the teacher model's encoder and decoder, respectively.

The objective function of the student model is,

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(x) &= \mathbb{E}_{\theta_\phi(z_c,z_d|x)}[\log P_\theta(x \mid z_c, z_d)] - \mathrm{KL}[Q_\phi(z_c, z_d \mid x) \mid\mid P(z_c, z_d)] \\ &\quad + (1-w)KL[Q_\phi(z_c, z_d \mid x) \mid\mid Q_\Phi(z_c, z_d \mid x)] \\ &\quad - \lambda I(\hat{x}, z_c)\end{aligned}$$

where, $x \sim p(w)P(x \mid w), \quad w \sim \mathrm{Ber}(\pi)$

Here for i-th dataset, $\pi = \frac{1}{i+1}$. Since student model will take data from both the synthetic data (generated by the teacher model) and current dataset, $x \sim p(w)P(x \mid w)$ represents how the student model will take the input data.

Some important expressions in the above objective function are,

- $\mathbb{E}_{\theta_\phi(z_c,z_d|x)}[\log P_\theta(x \mid z_c, z_d)] - \mathrm{KL}[Q_\phi(z_c, z_d \mid x) \mid\mid P(z_c, z_d)]$ is ELBO equation for general Variational auto encoder.
- $(1-w)KL[Q_\phi(z_c, z_d \mid x) \mid\mid Q_\Phi(z_c, z_d \mid x)]$ is used for posterior regularization which will rely on the fact that instead of doing KL divergence on full posterior, we can always change our posterior when we have given new information, that means posterior that we have used earlier will now become our prior.
- $I(\hat{X}, z_c)$ is used for incorporating mutual information which will ensure to enforce discriminate information about each distribution in model. They have achieved this by minimizing information in between decoded variable $\hat{x}$ and variable z_c, which they called information gain.

2.3 Lao, Qicheng et al. "FOCL: Feature-Oriented Continual Learning for Generative Models."

Earlier evaluation methods used in continual learning were basically focusing on quality of image (e.g.FID), however to measure how much model suffers from catastrophic forgetting, forgetfulness score meric was introduced by FOCL paper. Following forgetfulness score FS_t was used in paper:

$$FS_t = \frac{1}{t-1}\sum_{i=1}^{t-1}(d_t^{(i)} - d_i^{(i)}), t > 1 \tag{1}$$

Here $d_i^{(i)}$ is distance between true data distribution and generated data distribution of model for task 1 to $t-1$, similarly same distance d_t^i was recomputed for task t. Distance $(d_t^{(i)} - d_i^{(i)})$ basically infer how much there is forgetting from task i to current task t.

Overall forgetfulness score (weighted average of tasks) which was used is as following:

$$FS = \frac{2}{T * (T-1)} \sum_{t=2}^{T} (t-1) FS_t \tag{2}$$

3 Resources Utilized

The following resources were utilized:

- Software
  - Python=3.7.3
  - pillow=8.2.0
  - NumPy=1.20.1
  - SciPy=1.6.3
  - PyTorch=1.2.0
  - torchvision=0.9.1
  - opencv-python=4.5.1.48
  - visdom=0.1.8.9
  - torchfile=0.1.0
- Platform
  - Google Colab
  - IIT Kanpur CSE GPU: GeForce GTX 1080Ti 11GB memory

4 Experimental Results

4.1 GAN:

Figure 5 shows smooth transitioning from human faces to flowers. As can be seen, we can notice the subtle changes in illumination and contrast, followed by changes in structure, color and localization and finally refinements to the generation. Plot 6 shows the relation between FID scores and number of Iterations whilst training the GAN model and we can see the decrease in FID score which becomes constant later.

Figure 7 shows the generated images for the Imagenet Dataset by the model. The categories considered were fish, dog, butterfly, bird and snake. Table 1 contains the corresponding FID scores for the tasks learnt in the sequential order. As we can observe from the generated images and corresponding FID scores, we notice that some of the images fail to generate appropriately (Eg: Warping of fish images, squashing of dog images) and yet have a quite low FID score. This implies that the FID metric does not hold well in some scenarios, possibly in images which do not localize over the target well (with respect to the results acquired). Furthermore, we have used 1 to get a forgetful score metric (forgetfulness score of 0.473). Lower the score, the better) which facilitates in quantifying the forgetfulness of the model.

We experimented on few shot learning by varying the number of data samples on the flower dataset (1, 2, 4, 8, 16, 32, 64, 128, 512, 1024 samples).Plot 9 corresponds data samples ranging between 32 to 1024 and plot 9 corresponds to data samples ranging between 1 to 16. We noticed mode collapse in the data samples from 1 to 128, with generation quality of images maintained. On increasing the number of data samples from 128 to 512 we noticed a reduction in mode collapse and unfortunately a reduction in the quality of the image. From this, we infer that the model is going from one optima to another where the earlier optima suffered from mode collapse with good quality

generation and the current optima learnt over the 512 data samples shows less mode collapse with poor quality image generation. 1024 data samples have less mode collapse compared to 512 data samples and there is an improvement in the generated images quality.
Additionally, the findings are not quantified by the FID scores. Which is why we can also propose that the FID score does not always act as a good metric to quantify the mode collapse problem of the GAN and the quality of the generated image. The issues mentioned can be viewed in images 10 and 11.

4.2 VAE

For reproducing the results of continual learning in VAE, we are considering 4 tasks.

- Task1 - MNIST
- Task2 - regenerated data of Task1 + FashionMNIST
- Task3 - regenerated data of Task2 + SVHN
- Task4 - regenerated data of Task3 + CIFAR10

From fig. 16, 17 it can be seen that the generation capability of VAE degrades as the new task comes, especially the generation of color images are poor and some of the images are blacked-out. Because of the poor generation of images, the synthetically generated images by teacher model that are used to train the new student model are not good, and so the VAE becomes poor at reconstruction as the new task comes. And we from fig. 12,13,14,15 it is clear that both the ELBO and KL-divergence of the VAE are increasing as the new task comes. For continual learning of the above 4 task, VAE produces the forgetfulness score[3] as 8.59.

For performing few-shot learning the model was trained on FASHION-MNIST for 60000 sample and 256 batch-size for 60 epochs. The graph for FID-Score plotted for pre-trained model is in figure 18

After pre-training few short learning was performed on MNIST data-set. For performing the few shot learning a student model was initialized with MNIST data set and model obtained from pre-trained model was copied into teacher of this new MNIST data-set model. After that it was run for 300 epochs with sample size of 2, 4, 8, 16, 32, 64, 128, 512, 1024. Here sample-size means that we have randomly taken that many examples from train and test set both for performing few shot learning. FID score was computer on test-set.
Figure 19 shows the graph for FID-score calculate for each sample-size of 2 to 16 and 20 shows the graph for FID-score calculate for each sample-size of 32 to 1024 and From the graph it can be seen that, for less number of sample the FID score has zig-zag pattern because of the possibility of overshooting in model weights. And on other hand when we increase the sample size, graph stabilizes (we have noticed this trend in sample size 32 and larger sized samples). Also it is obvious that when we increase the sample size, the model learns well and so the FID-score is lower than the FID-scores of lesser samples in training.
Generated images for different number of data samples 8, 16, 32, 64, 128, 512, 1024 can be seen from 21, 22, 23, 24, 25, 26, 27 respectively. We can see from generated images that from sample size 32 on-words model has generated quite good quality images from epoch 100 and it further improves with more number of epochs.

5 Conclusion

In the previous sections we have reproduced the results of the literature considered over GAN and VAE, measured the "realistic" quality of the generated images and measured the forgetfulness of the two approaches via the forgetfulness metric mentioned in [3] which allowed us to compare the continual learning capabilities of the two approaches. Furthermore, we were able to generate images over a domain perceptually-distant from the domain the models have been trained over with relative small sample sizes. The above was performed with modifications (and a few reconstructions) over

the source code provided in the base literature. This led us to acquire the following through out the duration of the study,

- **Catastrophic Forgetting and Continual Learning**: The experimentation over the sequence of tasks over the GAN and VAE models clearly showcases the Catastrophic Forgetting problem that occurs in Neural Networks. On one hand with he VAE, we can clearly notice the Catastrophic Forgetting issue and this in turn allows us visualize the Continual Learning capabilities of the two approaches.
- **Visualized the Mode posterior problem of the GAN**: By varying the number of training sample sizes whilst modulating the source GAN towards a target domain, we we're clearly able to notice the Mode posterior problem within the results as discussed in the previous section.
- **Style parameters and smooth Interpolation**: From section 4.1, we saw the effect of the style parameters over the modulation of the source image. This follows the observations made by the base paper over the modulation flow (change in overall contrast and illumination, efforts in changing the style and further refining the details of the image) and the efficiency provided by normalizing the weights (i.e. removes the source style to help incorporate the target style).
- **Proves the universal structure**: Provided well localized images, the GAN approach was capable of showcasing the ability to modulate the source image to perceptually-distinct targets, which gives concreteness to the hypothesis that there could exist a common "blueprint" amongst images.
- **FID Scores and problems associated with it**: From section 4.1, we have observed the FID scores of the image generated by the GAN to be quite less given that the image generated does not represent the target domain appropriately (Eg: squashing of images generated over task dog).

6 Future Work

We have noticed that the generative-replay based approach utilized by the GAN holds promise in preventing catastrophic forgetting (and consequently maintains continual learning). Although the approach depends on saving the style parameters that would facilitate the modulation of the source GAN to generate target images, [1] have also leveraged techniques such as matrix factorization and low-rank regularization to compress the size of the style parameters. This approach can be imposed to the VAE model, where given a well trained VAE over a particular source, the latent vectors corresponding to the target for which the VAE generates images can be stored. This would prevent the VAE from suffering from catastrophic forgetting. A reason why it would be beneficial to impose GAN techniques over a VAE would be the fact that VAE do explicitly provides a sample distribution. Gaining higher quality images over the VAE model would enable us to gain information over the true distribution of the domain which could be beneficial. Furthermore, we have seen that the FID score although is a good metric for image quality does not always work well, which gives more scope for a better metric to assess the quality of an image.

Figures and Tables

Figure 5: Generated Images from GAN: Depicts the smooth interpolation from source to target (From Left to Right: iteration value 0, 1K, 2K, 59K, 60K)

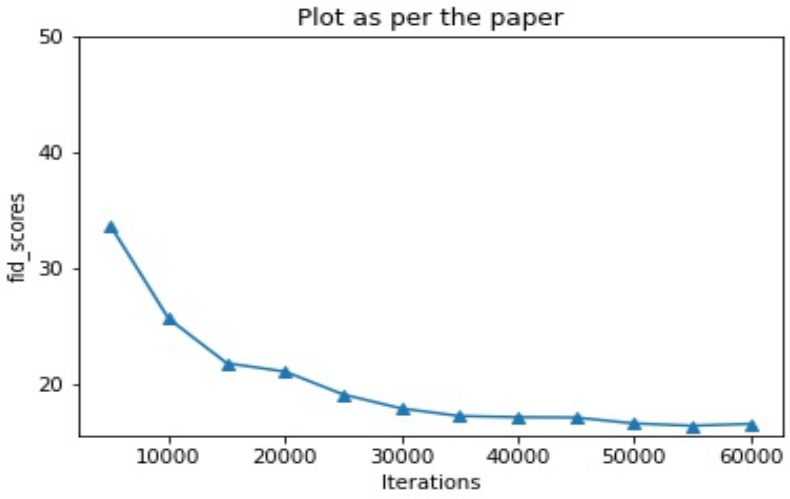


Figure 6: Relationship between the FID scores and Iterations as per paper [1]

Figure 7: Images generated by the model trained over Imagenet categories fish, dog, bird, butterfly, snake (Left to Right) as sequential tasks (Top to Bottom)

	Fish	**Dog**	**Bird**	**Butterfly**	**Snake**
Task 1	72.34				
Task 2	68.27	64.15			
Task 3	72.72	63.93	47.85		
Task 4	68.74	68.13	48.59	46.95	
Task 5	72.29	63.27	47.65	46.15	83.29

Table 1: FID scores while training for sequential tasks over GAN as shown in Figure 7

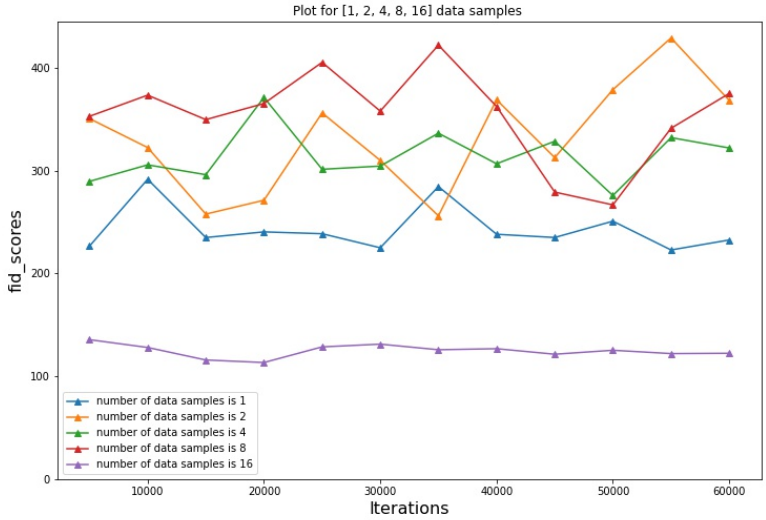


Figure 8: Relationship between the Fid scores and Iterations over sample size 1, 2, 4, 8, 16

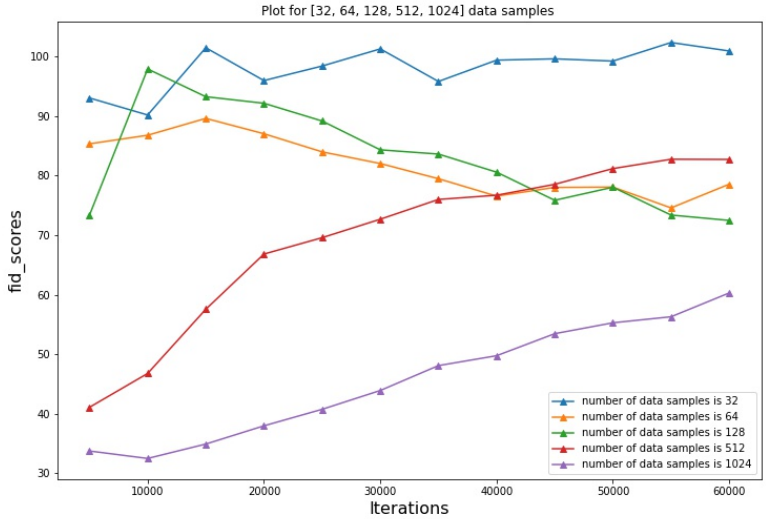


Figure 9: Relationship between the Fid scores and Iterations over sample size 32, 64, 128, 512, 1024

Figure 10: Generated Images for varying number of data samples (1, 2, 4, 8, 16, 32, 64, 128). Mode collapse lessens as number of samples increase.

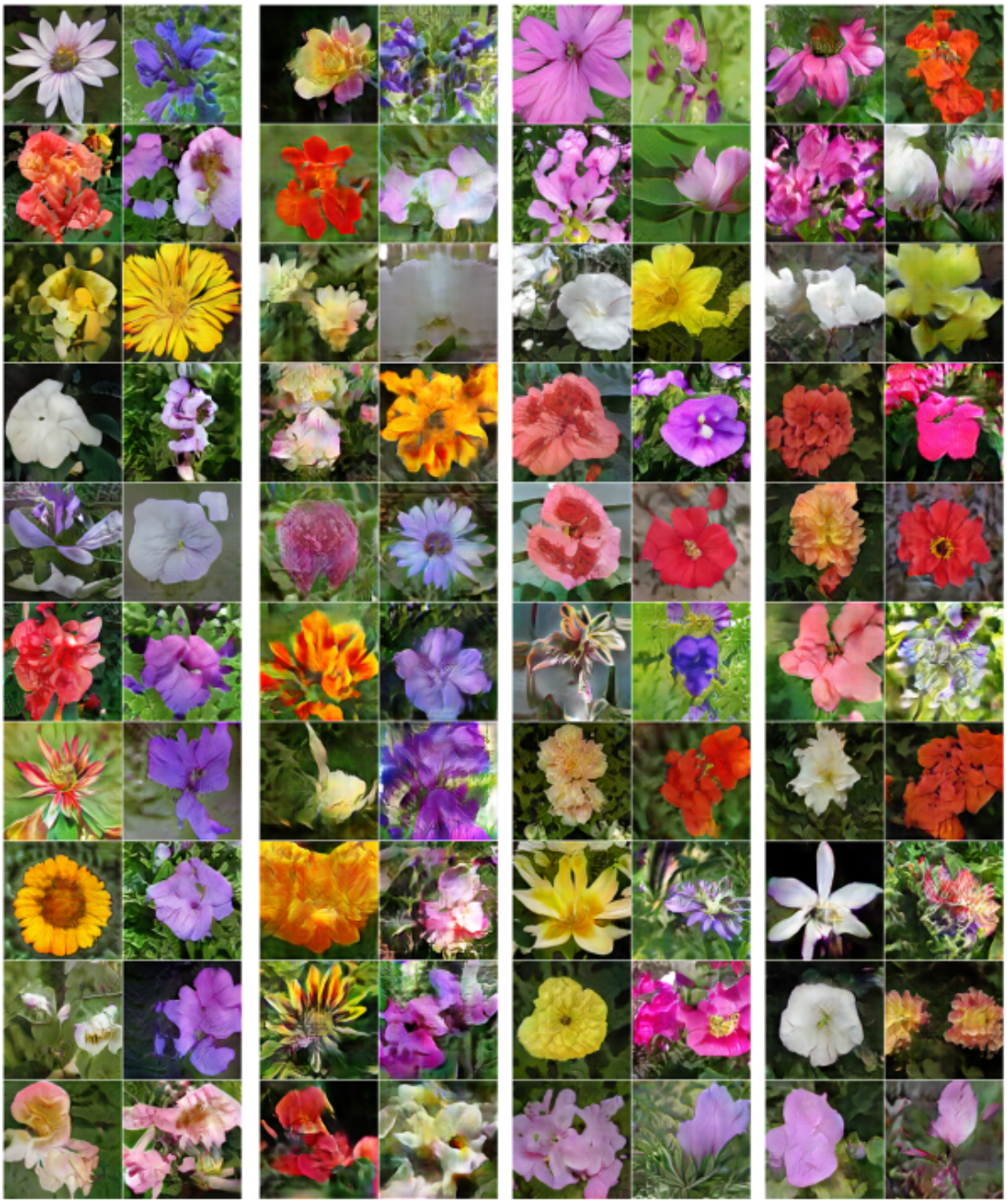


Figure 11: Generated Images for varying number of data samples (Left to Right: 512-iteration 5K, 512-iteration 55K, 1024-iteration 10K, 1024-iteration 60K). Mode collapse is not noticed. Blurriness noticed from 512-iteration 5K to 512-iteration 55K.

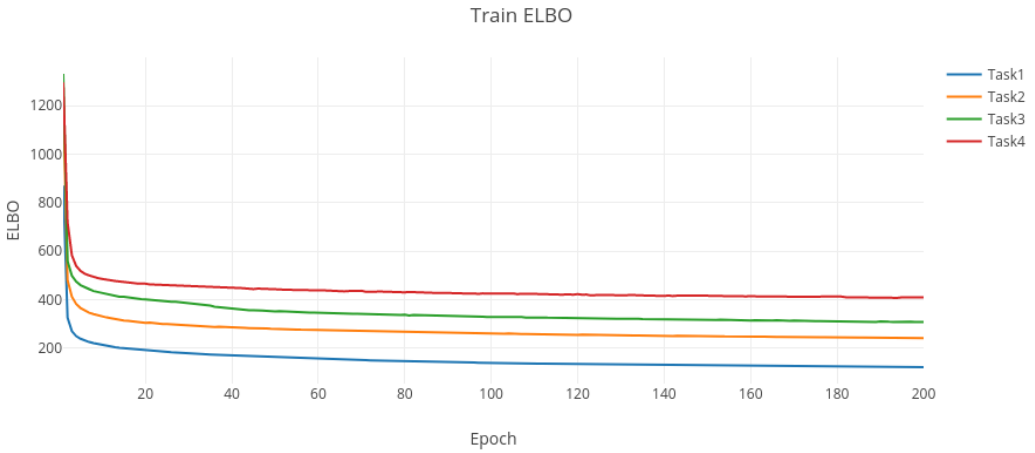


Figure 12: Train ELBO

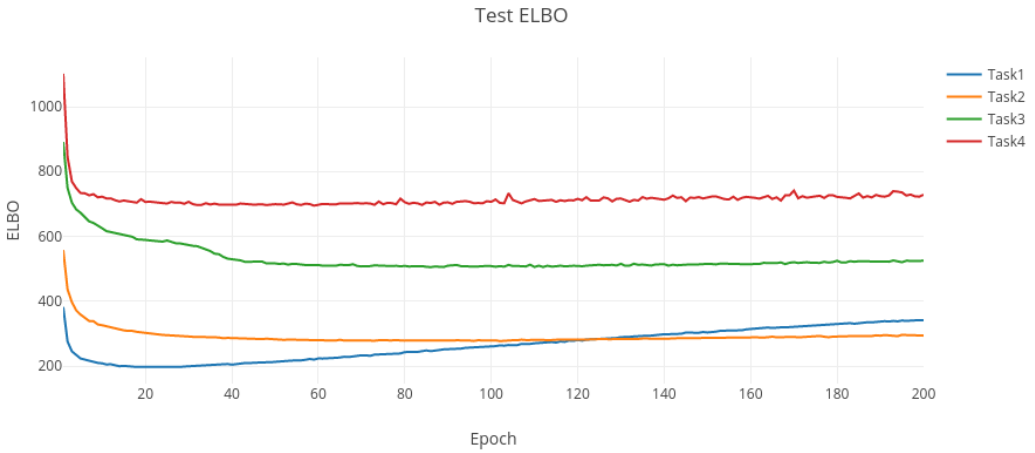


Figure 13: Test ELBO

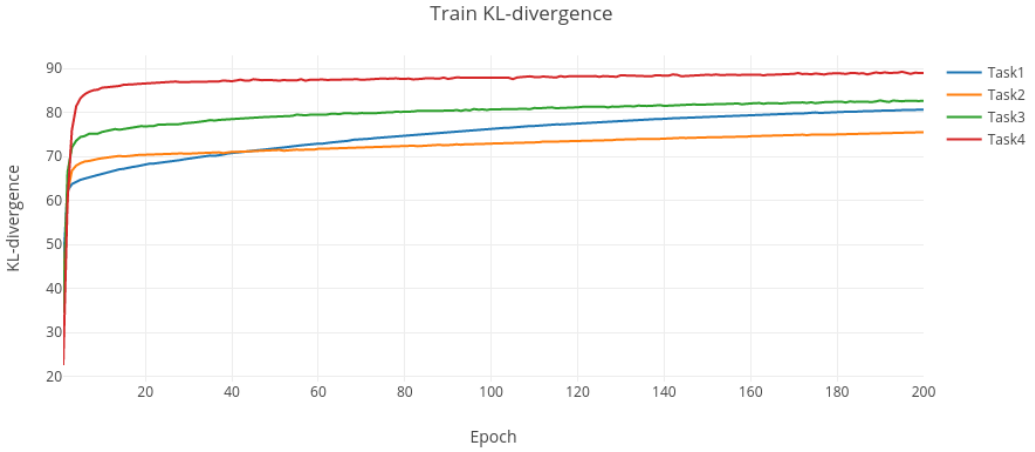


Figure 14: Train KL-divergence

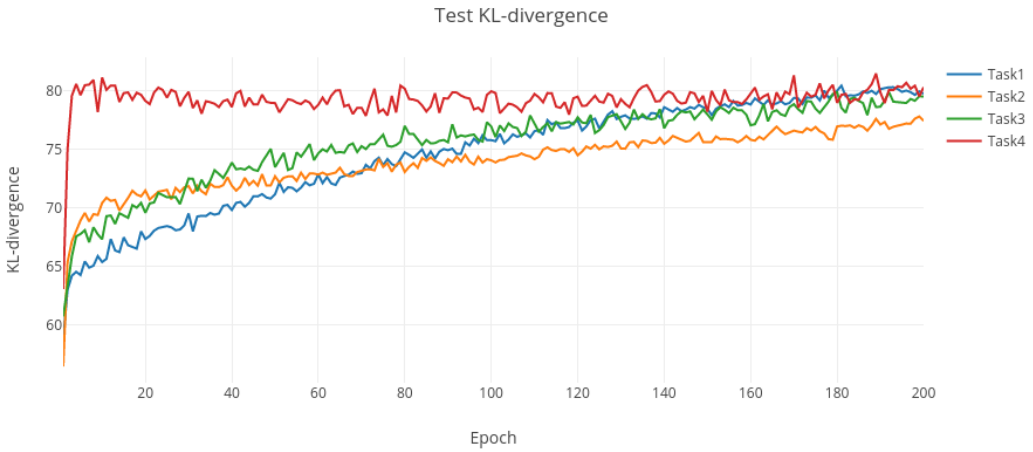
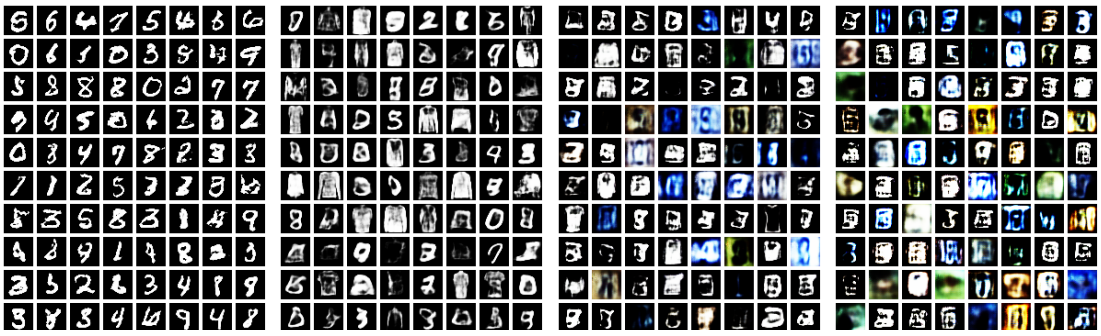
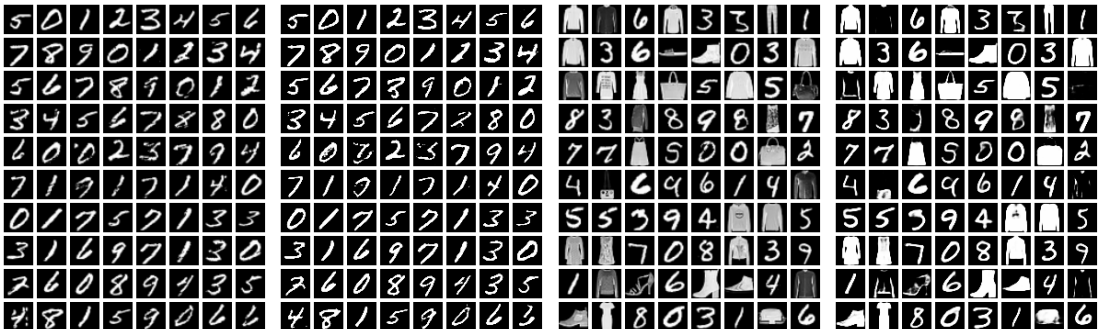


Figure 15: Test KL-divergence



(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4

Figure 16: Generation of Student VAE model



(a) Task 1: original (b) Task 1: reconstruction (c) Task 2: original (d) Task 2: reconstruction

(e) Task 3: original (f) Task 3: reconstruction (g) Task 4: original (h) Task 4: reconstruction

Figure 17: Test image and its reconstruction by Student VAE model

	MNIST	FashionMNIST	SVHN	CIFAR10
Task 1	172.61			
Task 2	178.33	149.74		
Task 3	175.87	172.72	188.46	
Task 4	185.81	204.12	214.34	220.11

Table 2: FID scores while continual learning for multiple tasks using VAE

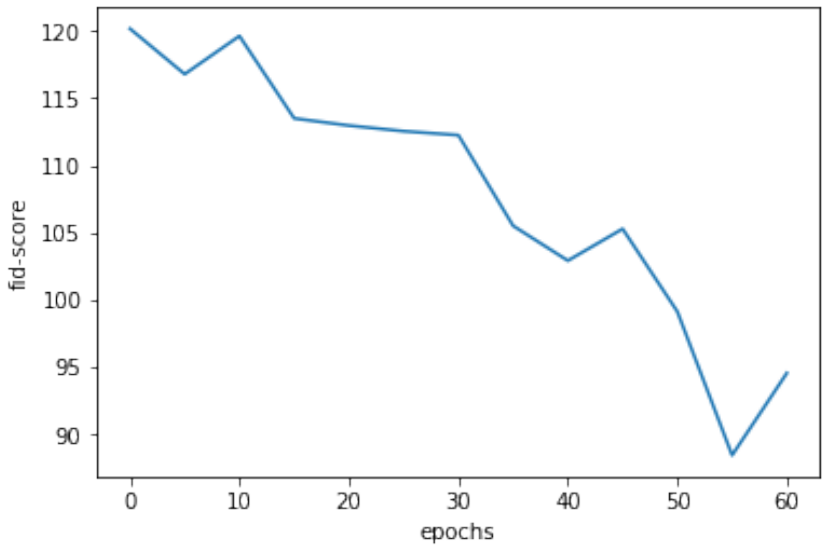


Figure 18: Fashion mnist-fid for batch size 256

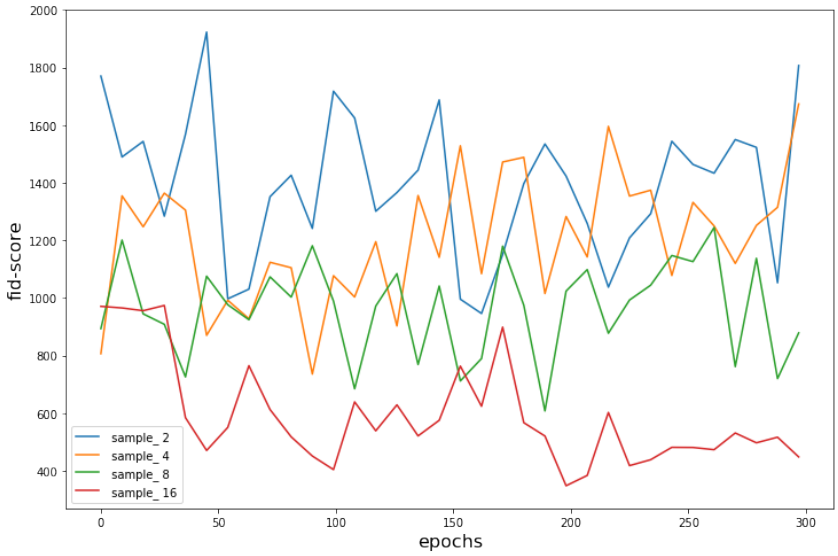


Figure 19: mnist fid plots for sample size 2 to 16

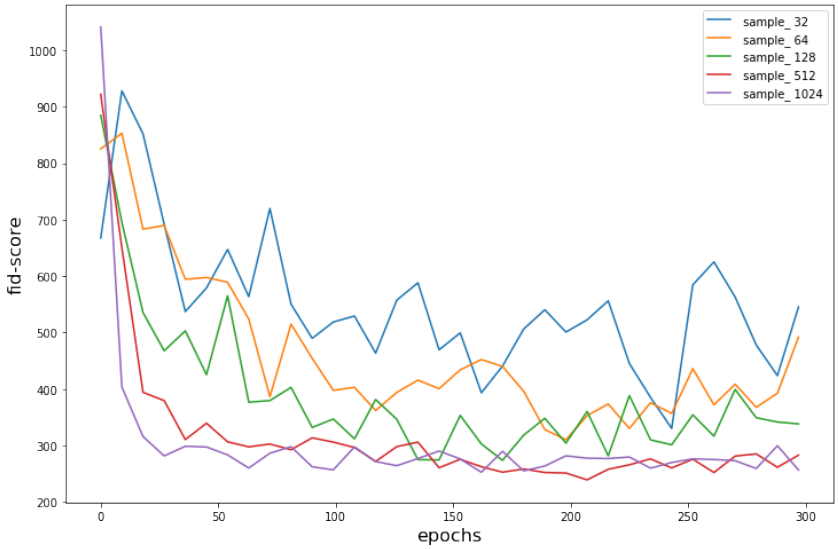


Figure 20: mnist fid plots for sample size 2 to 16

(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

Figure 21: VAE construction for sample size 8

(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

Figure 22: VAE construction for sample size 16

(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

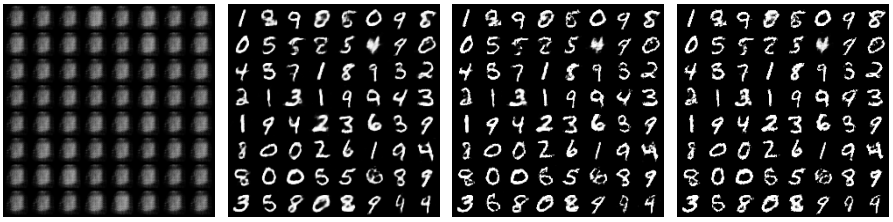
Figure 23: VAE construction for sample size 32

(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

Figure 24: VAE construction for sample size 64

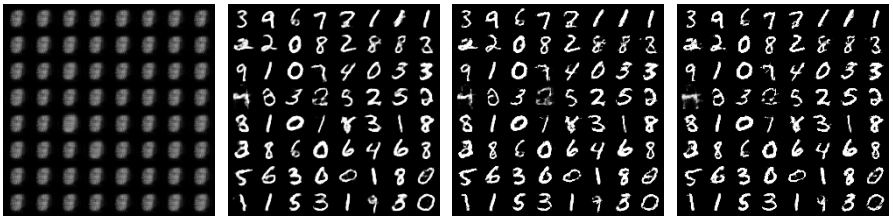
(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

Figure 25: VAE construction for sample size 128



(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

Figure 26: VAE construction for sample size 512



(a) # epochs 1 (b) # epochs 100 (c) # epochs 200 (d) # epochs 300

Figure 27: VAE construction for sample size 1024

References

[1] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. *arXiv preprint arXiv:2006.07543*, 2020.

[2] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[3] Qicheng Lao, Mehrzad Mortazavi, Marzieh Tahaei, Francis Dutil, Thomas Fevens, and Mohammad Havaei. Focl: Feature-oriented continual learning for generative models. *arXiv preprint arXiv:2003.03877*, 2020.

[4] Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Lifelong generative modeling. *arXiv preprint arXiv:1705.09847*, 2017.