**Program Assignment 4**

**Deep Generative Modeling: Diffusion Models**

**Problem Set for DDIM**

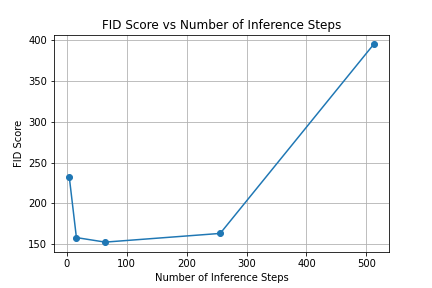
**Question 1: Using the LSUN Church dataset and change the number of steps, analyze how the number of steps affects the quality of generated images. Please use both quantitative measure (FID score) and qualitative measure (a series of generate images at different settings of timesteps) to report your findings.**

To investigate the qualitative differences between the generated images and the original images under varying diffusion steps, we first set the diffusion steps to 4, 16, 64, 256, and 512. Using these diffusion steps, we generated images for ten different samples. The resulting images are visually aligned with their corresponding original images for a straightforward comparison. For this experiment, we set the parameter η to 0, ensuring that we are operating under the standard DDIM framework.

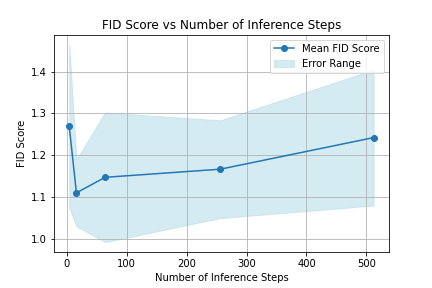
This approach allows us to explore the impact of different diffusion steps on the fidelity and quality of the generated images relative to the original images.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Original figure | T = 4 | T = 16 | T = 64 | T = 256 | T = 512 |
| DDPM_image_1 |  | DDIM_image_1 | DDIM_image_1 | DDIM_image_1 | DDIM_image_1 |
| DDPM_image_2 | DDIM_image_2 | DDIM_image_2 | DDIM_image_2 | DDIM_image_2 | DDIM_image_2 |
| DDPM_image_3 | DDIM_image_3 | DDIM_image_3 | DDIM_image_3 | DDIM_image_3 | DDIM_image_3 |
| DDPM_image_4 | DDIM_image_4 | DDIM_image_4 | DDIM_image_4 | DDIM_image_4 | DDIM_image_4 |
| DDPM_image_5 | DDIM_image_5 | DDIM_image_5 | DDIM_image_5 | DDIM_image_5 | DDIM_image_5 |
| DDPM_image_6 | DDIM_image_6 | DDIM_image_6 | DDIM_image_6 | DDIM_image_6 | DDIM_image_6 |
| DDPM_image_7 | DDIM_image_7 | DDIM_image_7 | DDIM_image_7 | DDIM_image_7 | DDIM_image_7 |
| DDPM_image_8 | DDIM_image_8 | DDIM_image_8 | DDIM_image_8 | DDIM_image_8 | DDIM_image_8 |
| DDPM_image_9 | DDIM_image_9 | DDIM_image_9 | DDIM_image_9 | DDIM_image_9 | DDIM_image_9 |
| DDPM_image_10 | DDIM_image_10 | DDIM_image_10 | DDIM_image_10 | DDIM_image_10 | DDIM_image_10 |

Subsequently, we calculated the FID (Fréchet Inception Distance) values by comparing the original images with the generated images under different diffusion steps, using a set of 10 images as the dataset. The computed FID values provide a quantitative measure of the differences, which are presented in the following graph.



Subsequently, we plotted the FID indices for each pair of corresponding images. Since FID is calculated for a set of images, we applied data augmentation techniques and computed the FID values for each pair among the ten images. The resulting data were used to plot error bars and mean values, as shown in the figure below. This approach resulted in a relatively smoother data representation.



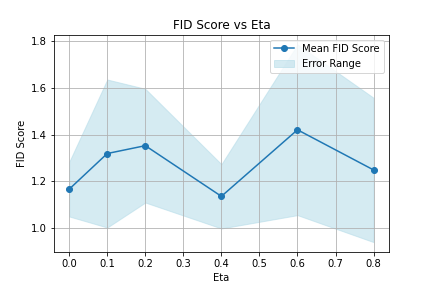
As depicted in the image, the generated outputs with fewer diffusion steps (T=4 and T=16) show significant deviation from the originals, displaying high levels of blur and noise. However, as the number of diffusion steps increases (T=64 and T=256), the generated images exhibit progressively better quality and closer resemblance to the original images. At T=512, the images tend to become overly detailed with noise, suggesting an overfitting effect or an excessive level of detail introduced by too many diffusion steps. These observations underline the delicate balance required in choosing the appropriate number of diffusion steps to optimize the quality of image generation while maintaining computational efficiency.

**Question 2 (4 points): As discussed before, η is an essential hyper-parameter in DDIM that controls the level of stochasticity during the reverse denoisng process. By adjusting η, you can balance between determinism (for faster sampling with fewer steps) and stochasticity (for greater sample diversity).**

**Please use both quantitative measure (FID score and sampling time in hours/minutes) and qualitative measure (a series of generate images at different settings of η) to report your findings.**

Subsequently, we sought to explore the relationship between different values of η and the resulting generated images, essentially examining the effect of randomness on image generation. In this investigation, we selected the inference step count of 256, which demonstrated the best performance in the previous task. We then analyzed the variations in generated images under different levels of randomness η. The results are illustrated in the figure below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Original figure | Ƞ = 0.2 | Ƞ = 0.4 | Ƞ = 0.6 | Ƞ = 0.8 | Ƞ = 1.0 |
| DDPM_image_1 | DDIM_image_1 | DDIM_image_1 | DDIM_image_1 | DDIM_image_1 | DDIM_image_1 |
| DDPM_image_2 | DDIM_image_2 | DDIM_image_2 | DDIM_image_2 | DDIM_image_2 | DDIM_image_2 |
| DDPM_image_3 | DDIM_image_3 | DDIM_image_3 | DDIM_image_3 | DDIM_image_3 | DDIM_image_3 |
| DDPM_image_4 | DDIM_image_4 | DDIM_image_4 | DDIM_image_4 | DDIM_image_4 | DDIM_image_4 |
| DDPM_image_5 | DDIM_image_5 | DDIM_image_5 | DDIM_image_5 | DDIM_image_5 | DDIM_image_5 |
| DDPM_image_6 | DDIM_image_6 | DDIM_image_6 | DDIM_image_6 | DDIM_image_6 | DDIM_image_6 |
| DDPM_image_7 | DDIM_image_7 | DDIM_image_7 | DDIM_image_7 | DDIM_image_7 | DDIM_image_7 |
| DDPM_image_8 | DDIM_image_8 | DDIM_image_8 | DDIM_image_8 | DDIM_image_8 | DDIM_image_8 |
| DDPM_image_9 | DDIM_image_9 | DDIM_image_9 | DDIM_image_9 | DDIM_image_9 | DDIM_image_9 |
| DDPM_image_10 | DDIM_image_10 | DDIM_image_10 | DDIM_image_10 | DDIM_image_10 | DDIM_image_10 |



In comparing and evaluating the results, we also assessed the pairwise differences between the generated images and the original images for each value of η. Given that FID compares results between two sets of images, we applied data augmentation techniques such as rotation, translation, and stretching to the same image. This approach allowed for a more robust analysis of the results. The findings were then analyzed, and the error bars were plotted accordingly.

Based on the qualitative and quantitative results, we can draw the following conclusion:

The image quality does not significantly correlate with the randomness parameter

η; however, results tend to be relatively better when η is at intermediate values (0.5).

**Question 3 (3 points):**

1. **What does the word implicit actually mean in DDIM?**
2. **What do you think are some limitations of DDIM?**

**Please answer this question in text format. Note that this is an open question, and there are no right or wrong answers.**

In DDIM (Denoising Diffusion Implicit Models), the word "implicit" refers to how the model generates samples in a less direct way. Instead of using a clearly defined random process, DDIM uses a series of steps that gradually clean up noise to create realistic data samples.

Here's a breakdown:

**Denoising Steps:** DDIM works by starting with noisy data and then taking multiple steps to remove the noise bit by bit. This process eventually turns the noisy data into something that looks like real data.

**Implicit Method:** Traditional models use explicit random processes to generate samples. In contrast, DDIM uses a deterministic (predictable and non-random) function to convert noise into data. This is what makes it "implicit."

**Predictable Sampling:** DDIM can produce the same result every time by tweaking the denoising steps, making the process more stable and predictable.

Therefore, "implicit" in DDIM means it uses a more indirect and predictable method to turn noise into data, rather than relying on clear-cut random processes. This makes DDIM more stable and controllable when generating samples.

(ii)Based on the mechanism of DDIM and the experiments provided in the images, several limitations of DDIM can be identified:

**Quality Degradation at High Diffusion Steps:** As seen in the images, when the number of diffusion steps (T) is very high (e.g., T=512), the generated images tend to become overly detailed with noise, leading to an overfitting effect or excessive levels of detail. This results in poor image quality and a high FID score.

**Blur and Noise at Low Diffusion Steps:** With fewer diffusion steps (e.g., T=4 and T=16), the generated images show significant deviation from the originals, displaying high levels of blur and noise. This indicates that DDIM may not be effective in capturing fine details of the original images when the number of steps is too low.

**Sensitivity to Hyperparameters:** The results indicate that the quality of generated images is sensitive to the hyperparameter settings, particularly the number of diffusion steps and the parameter η. Finding the optimal settings requires extensive experimentation, which can be computationally expensive and time-consuming.

In conclusion, while DDIM can generate high-quality images under certain conditions, its e ffectiveness is highly dependent on the appropriate selection of diffusion steps and hyperparameter settings. The model's sensitivity to these settings and the potential for quality degradation at extreme values are notable limitations that require careful consideration and experimentation.

**Question 2: Modify the number of epochs and discuss how results change in different epoch settings.**

We first set the parameters manual\_Seed = 999, lr = 0.0002, batch\_size = 128. Under these consistent conditions, we observe the changes in the data at epochs = 1, 2, and 3.

|  |  |  |
| --- | --- | --- |
| Epochs =1 |  | **Average G\_loss**  **5.19535**  **Average D\_loss**  **0.64196** |
| Epochs =2 |  | **Average G\_loss**  **4.29896**  **Average D\_loss**  **0.67143** |
| Epochs =3 |  | **Average G\_loss**  **3.85396**  **Average D\_loss**  **0.68477** |

From the results above, we can see that as the epochs increase, G's loss gradually decreases, while D's loss tends to increase. However, from the images, it is evident that the quality of the generated pictures becomes more refined, but they do not accurately reproduce the original details of the images (such as hair, clothing, etc.) Since the objective of the GAN loss function is to minimize G's loss and maximize D's loss, from the perspective of the loss function, the model's error decreases as the number of epochs increases.

**Question 3: Change the learning rate and batch size and discuss the effects of these hyperparameters on the training process.**

First, we explore the impact of the learning rate on training, maintaining model parameters manualSeed = 999, num\_epochs = 2, batch\_size = 256, while adjusting the lr values to 0.0001, 0.0005, 0.001 and observing the results, as shown below:

|  |  |  |
| --- | --- | --- |
| Learning rate 0.0001 | real_vs_fake | **Average G\_loss**  **8.63111**  **Average D\_loss**  **0.28469** |
| Learning rate 0.0005 | real_vs_fake | **Average G\_loss**  **3.24576**  **Average D\_loss**  **1.16500** |
| Learning rate 0.001 | real_vs_fake | **Average G\_loss**  **3.16951**  **Average D\_loss**  **1.10190** |

From the figure above, we can see that when the learning rate is set to 0.0001, the images remain relatively blurry after iterative training, particularly with facial details being difficult to discern. Essentially, all images exhibit similar features with little variation. This may be attributed to the small learning rate causing minimal parameter updates, as reflected by the average values of the loss function. On the other hand, when the learning rate is set to 0.001, although the loss function indicates smaller errors compared to a rate of 0.0005, the generated images incorporate many unnecessary noise points, such as pink dots in the background. Therefore, overall, setting the learning rate to 0.0005 yields the best performance of the model.

Next, we explore how changes in batch size affect the generation of images. In this case, we assume manual\_Seed = 999, num\_epochs = 2, and lr = 0.0003, with batch sizes of 128, 256, 512 , 1024for variation. The results are as follows:

|  |  |  |
| --- | --- | --- |
| **Batch Size 128** | **real_vs_fake** | **Average G\_loss**  **3.77979**  **Average D\_loss**  **0.93408** |
| **Batch Size**  **256** |  | **Average G\_loss**  **4.24599**  **Average D\_loss**  **0.89057** |
| **Batch Size**  **512** | real_vs_fake | **Average G\_loss**  **4.63432**  **Average D\_loss**  **0.99142** |
| **Batch Size**  **1024** | real_vs_fake | **Average G\_loss**  **6.43754**  **Average D\_loss**  **1.01046** |

From the figure above, we can see that as the batch size increases, particularly when it reaches 512 and 1024, the loss function significantly increases. This indicates that the model's bias also grows with the increase in batch size. This can be observed in the generated images as well; as the batch size increases, the details of the facial features and skin texture in the generated images deviate more from the original data. Therefore, we can conclude that the larger the batch size, the less precise the images generated by the GAN become.

**Bonus Question: Use the CIFAR-10 dataset to train the DCGAN model, analyze the results, and compare the performance of DCGAN on the Celeb-A Faces dataset and the MNIST dataset.**

Compile the results of the three datasets into the table below, as shown in the following figure.

|  |  |  |  |
| --- | --- | --- | --- |
| **CIFAR-10** | **manualSeed 999**  **batch\_size 128**  **num\_epochs**  **2**  **Learn Rate**  **0.00075** |  | **Average G\_loss**  **5.44961**  **Average D\_loss**  **1.05899** |
| **MINIST** | **manualSeed 999**  **batch\_size 128**  **num\_epochs**  **2**  **Learn Rate**  **0.00075** | **real_vs_fake** | **Average G\_loss**  **2.44961**  **Average D\_loss**  **1.50989** |
| **Celeb-A Faces** | **manualSeed 999**  **batch\_size 128**  **num\_epochs**  **2**  **Learn Rate**  **0.00075** |  | **Average G\_loss**  **2.53015**  **Average D\_loss**  **1.21244** |

Considering the images and the results of the loss function, we can roughly assume that, in terms of the DCGAN model, the performance on the MINIST and Celeb-A Faces datasets is superior to that on CIFAR-10. This is because the basic shapes of objects are not discernible in CIFAR-10, while they are apparent in the other two datasets. However, for the MINIST dataset, although the shapes generated are somewhat similar due to its simplicity, the numbers generated are not accurate (for example, a 0 is generated as a 7), indicating that there is room for improvement in the model’s accuracy.