Image Generation Using Diffusion Models

CREATIVE COMPONENT REPORT

Department Of Computer Science

OKLAHOMA STATE UNIVERSITY, STILLWATER

MAY 2023

SUBMITTED BY:
Srikar Amara
Banner Id: A20340640

Introduction

Diffusion models are a type of generative model that simulate random diffusion processes in the latent space to create new images. These models learn to continuously add Gaussian noise to the training data and then reverse this noise process to recover the original data. They are commonly used for image denoising, smoothing, and upscaling. In a diffusion model, the image generation process begins with a random point in the latent space, followed by a random walk that generates new points. These points are then transformed into images using a decoder network. The random walk can be controlled by adjusting the diffusion time, step size, or diffusion matrix to achieve the desired properties of the generated images.

Compared to other generative models, such as Variational Auto-encoders (VAEs) or Generative Adversarial Networks (GANs), diffusion models offer a flexible approach to image generation, as the diffusion process can be easily modified to achieve the desired results. Diffusion models have several uses, including content creation, data augmentation, gaming, augmented reality, and education.

Diffusion models are inspired by non-equilibrium thermodynamics and define a Markov chain of diffusion steps to slowly add random noise to the data. The model then learns to reverse the diffusion process to create desired data samples from the noise.

GANs, on the other hand, are a class of machine learning frameworks in generative AI. They consist of two neural networks, a generator and a discriminator, that play a game. The generator tries to fool the discriminator by generating data similar to that in the training set, while the discriminator tries to identify fake data from real data. GANs can be used to generate new examples of datasets, such as medical imaging, satellite imagery, and natural language processing. Game developers also use GANs to create new characters in video games and to generate audio and characters for non-playable characters (NPCs).

There are several methods for generating new images, including Generative Adversarial Networks and Variational Auto Encoders. The primary objective of this project is to build a GAN model using diffusion models to generate new images and evaluate its performance and limitations against other generative models, such as VAEs. With this analysis, researchers can make informed decisions on which model to use for specific training needs.

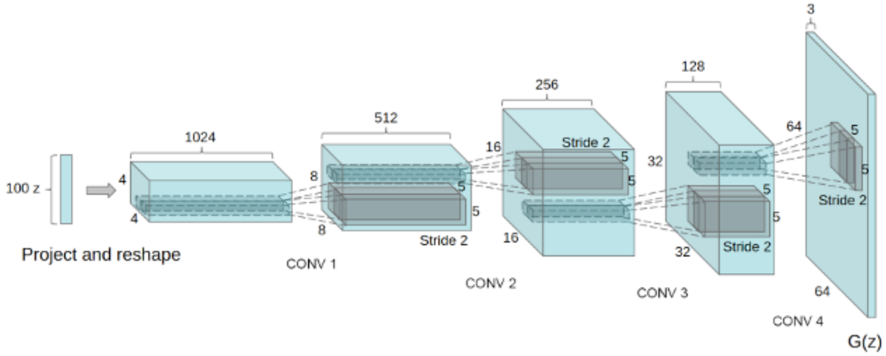

Figure 1. Generator neural network architecture.[5]

Background and Literature Review

Unsupervised representation learning has been a longstanding problem in computer vision, particularly in the context of images. The emergence of GANs as a promising approach in generative AI has led to extensive research into their applications in various domains. The introduction of GANs in 2014 by Ian Goodfellow sparked significant interest in their potential for unsupervised feature learning in natural images, as demonstrated by Li et al. (2017). This work showed that GANs can learn powerful image representations without relying on explicit supervision, making them a promising tool for unsupervised representation learning.

In addition to natural images, GANs have also been applied to image-to-image translation tasks, where they learn to map one type of image to another. For example, Zhu et al. (2017) introduced CycleGAN, a GAN-based model that can translate images between different domains (e.g., horses to zebras, or summer to winter). This work demonstrated that GANs can be used for unsupervised domain adaptation, without the need for paired training data.

GANs have also shown promise in voice generation, as demonstrated by Donahue et al. (2018), who introduced a GAN-based model for voice conversion. This work showed that GANs can be used to learn complex mappings between speech signals, allowing for high-quality voice conversion without the need for explicit alignment or supervision.

Overall, the extensive research on GANs highlights their effectiveness for unsupervised representation learning in various domains, and their potential for generating high-quality samples with diverse applications.

Technical Approach and methodology

Dataset Collection:

- Collected MNIST dataset, CelebA dataset from kaggle and random images dataset (64X64) from imagenet.

GAN Model Architecture:

- Implemented GAN architecture using Generator and Discriminator for MNIST dataset.
- Trained the GAN architecture with 5 epochs, 10 epochs and 15 epochs to find the generation capabilities of GAN architecture with multiple kinds and levels of datasets.
- Modified the architecture to work with CelebA dataset and ran the experiment to monitor the performance metrics for over 100 epochs.
- Calculated the efficiency of Generator and discriminator using BCE loss function and FID score for MNIST dataset and CelebA dataset.

VAE Model Architecture:

- Implemented the VAE architecture with an encoder with two convolutional layers followed by a fully connected layer and a decoder network with a mirrored architecture of encoder using a fully connected layer followed by three convolutional transpose layers.
- Trained the model to work with MNIST dataset for 5 epochs and 1 epoch.
- Modified the architecture to work with CelebA dataset and ran the experiment to monitor the performance metrics.
- Calculated the efficiency of VAE architecture using ELBO loss function
-

Overall, the project involved implementing and modifying GAN and VAE architectures, training the models on different datasets, and comparing the results. The project aimed to gain insights into the effectiveness and limitations of these models for unsupervised representation learning and image generation tasks on simple to complex datasets.

Results and Limitations

In comparison, VAE is easier to train than GAN for simpler datasets. But GAN's tend to generate high quality images when compared to VANs. For simpler dataset such as MNIST, VAE performed good when compared to GAN. VAE generated better images than GAN for just 5 epochs where GAN had to train for over 50 epochs to catch up to VAE.

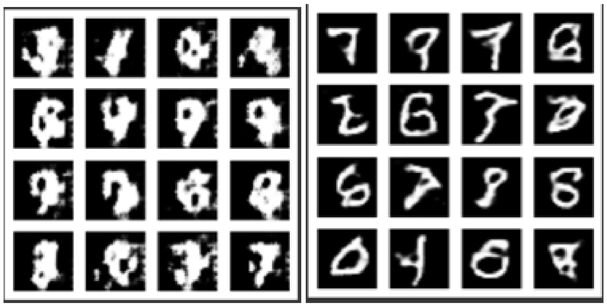

Figure 2. MNIST dataset for 10 and 50 epochs using GAN

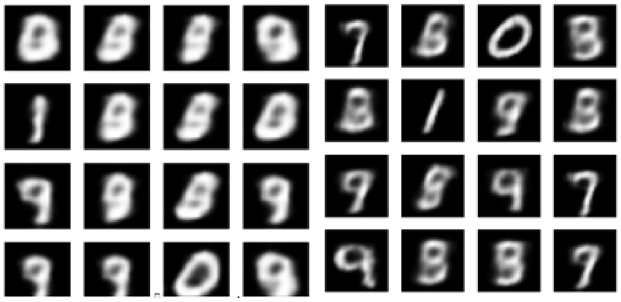

Figure 3. MNIST dataset for 1 and 5 epochs using VAE

For complex datasets such as CelebA dataset, GAN performed better when compared to VAE. GAN started to produce fake images for 10 epochs whereas VAE's generated blurry images even after trained for over 50 epochs as shown in fig :

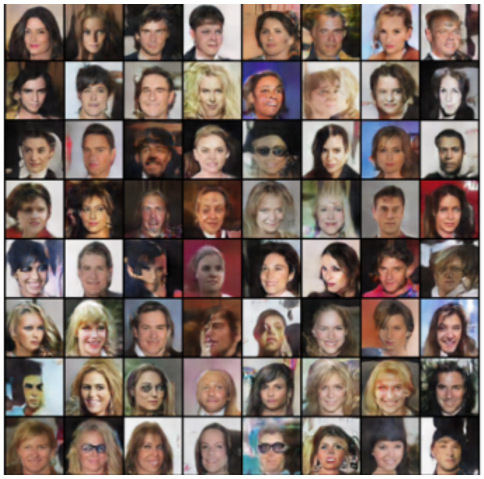

Figure 4.CELEBA dataset for 10 epochs using GAN

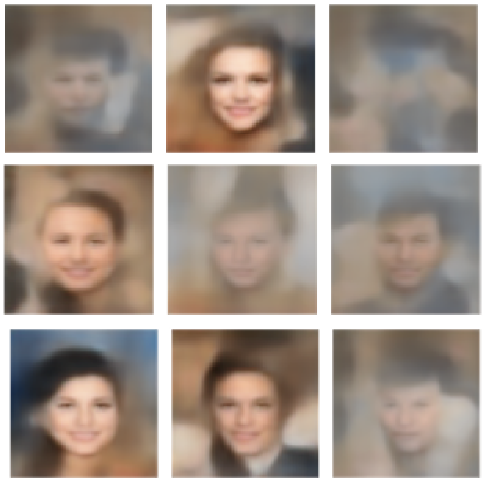

Figure 5.CELEBA dataset for 50 epochs using VAE

When GAN is trained for random images without specific context such as (Imagenet 64X64) dataset, GAN couldn't be able to generate images even after training the model for 100 to 150 epochs. The loss function is very high and didn't go down. This is due to the complexity of images. The GAN couldn't be able to distinguish the multiple images.

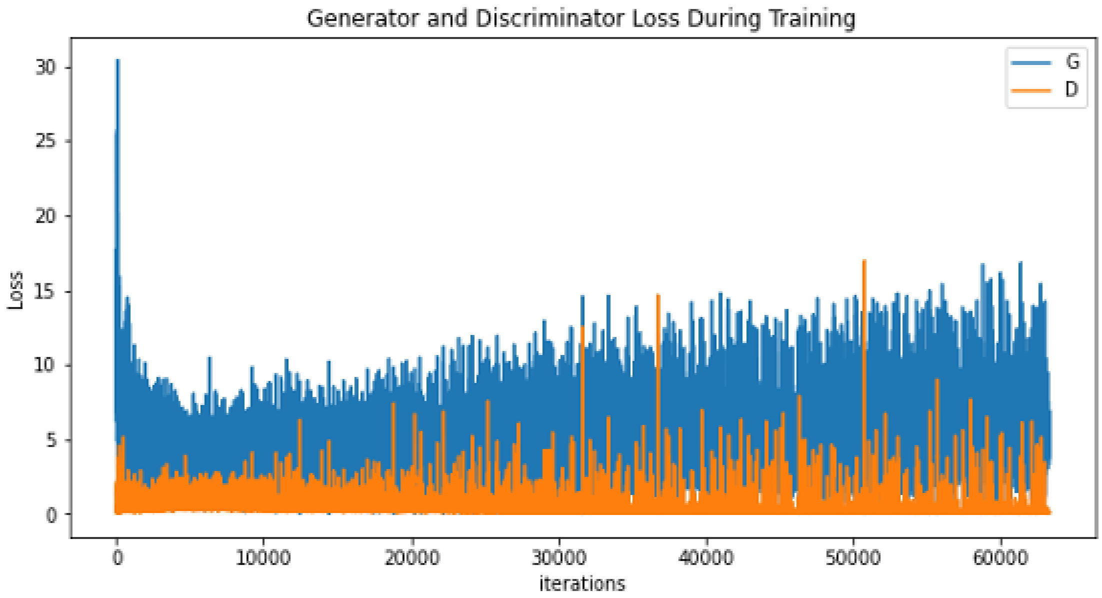

Figure 6. Generator and Discriminator Loss During Training for GAN

Loss values :

Loss values are calculated to check the performance of model. Here, GAN is using BCE loss whereas VAE is using ELBO loss. The lower the loss value, higher the accuracy for both BCE and ELBO loss values. The generated images for GAN and VAE are shown in Fig: 2. Fig: 3. Fig: 4. Fig: 5.

To measure the performance metrics for GAN, FID score is generated. FID (frechet inception distance) score measures the distance between the distributions of the generated images and the real images in the training set. A lower FID score indicates that the generated images are more similar to the real images, while a higher FID score indicates that the generated images are less similar to the real images.

For the MNIST dataset, the FID score is 175.0 for GAN. and for the CelebA dataset, the observed FID score is 120.23. The reason for this observation of FID score is that, for MNIST dataset, the images are relatively simple. And the MNIST dataset train data is lightweight. And the trainset for the CelebA dataset is heavy(i.e more samples to train) which could be some of the reasons the CelebA trained output is better compared to MNIST dataset.

To measure the performance of VAE, ELBO loss function is used. The lower the loss, higher the quality of the reconstructed image and higher the loss, lesser the quality of the image. The observed score for MNIST score for VAE is -157.4 and for the CelebA dataset, the score is 6327.64. For more complex datasets such as 64X64 Imagenet dataset, more complex generator and discriminator with more layers would stop training due to the GPU memory restrictions. The loss value comparison is shown in table 1.

Model Name	Training Dataset	Evaluation Dataset	Loss Function	ELBO Value For VAE	FID Score for GAN
GAN	MNIST	MNIST	BCE loss	-	175.0
VAE	MNIST	MNIST	ELBO	-157.4	-
GAN	CELEB FACE	CELEB FACE	BCE loss	-	120.23
VAE	CELEB FACE	CELEB FACE	ELBO	6327.64	-

Table 1. Evaluation Table For MNIST and Celebrity Face Datasets

Conclusion And Future Work

VAE metric is compared by ELBO loss function whereas GAN metric is compared by several factors such as FID score, Inception Score (IS), precision and recall and Intra-Class Distance (ICD). The lower the ELBO loss function, higher the accuracy and lower the FID score, higher the quality of the image.

This analysis helps researchers and students in making good decisions on which model to use for specific training needs. GAN can produce high quality images whereas there is a lot of room for future research. Such as generating high quality images for smaller datasets and working on Generator and discriminator stability. Diffusion models are used in other research areas such as image enhancement, audio generation and these diffusion models can be enhanced by focusing on stability of diffusion models (stable diffusion) as stable diffusion is often limited by functionality and image quality.

Future work can focus on combining the strengths of VAEs and GANs to develop hybrid models such as V-GANs. V-GANs can potentially improve the performance of both VAEs and GANs by leveraging the strengths of both models. For example, a V-GAN could use the encoder from a VAE to generate a latent representation, which could then be passed to the generator of a GAN to produce high-quality, realistic images.

In addition, future research could focus on developing new loss functions that are better suited for specific applications. For example, for image denoising, a loss function that penalizes the difference between the denoised image and the ground truth image could be developed. Similarly, for image segmentation, a loss function that encourages the generated images to have clear object boundaries could be developed.

Finally, the training process for both VAEs and GANs can be further optimized. This could involve exploring new optimization algorithms, adjusting hyperparameters, and developing new regularization techniques to prevent overfitting. Additionally, techniques such as curriculum learning and transfer learning could be explored to improve the training process for both VAEs and GANs.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.

[2] C. Li, M. Wand, J. Zhu, and X. Liu, "Triple generative adversarial nets," in Advances in Neural Information Processing Systems, 2017, pp. 4058-4068.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223-2232.

[4] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in Advances in Neural Information Processing Systems, 2018, pp. 6696-6705.

[5] https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf