Comparative Analysis of GAN and Diffusion Models for Text-to-Image Generation of Flower and Cub Images

Authors: Rajib Das
Emails: rajibloc@buffalo.edu
UB Person ID: 50546929

Abstract:

This report details a comparative study of Generative Adversarial Networks (GANs) and Diffusion models, emphasizing their application in generating images from textual descriptions of flowers and buds. This project leverages the advanced technique of creating a cross-embedding space between textual and visual data, facilitating the synthesis of detailed and contextually accurate images from text. The models are evaluated based on Inception Score (IS) and Fréchet Inception Distance (FID), alongside a qualitative assessment of the visual results. This work tests the models' capabilities in handling the complexity of bridging text and image modalities.

1. Introduction

The ability to generate images from textual descriptions bridges the gap between natural language processing and computer vision, presenting a compelling challenge in the field of artificial intelligence. This project explores this intersection by comparing the capabilities of Generative Adversarial Networks (GANs) and Diffusion models in creating detailed images of flowers and buds from text. Focusing on floral imagery, which demands high fidelity in color, shape, and texture representation, this study evaluates the models' ability to translate textual descriptions into corresponding visual outputs. By leveraging advanced generative techniques, the project aims to advance our understanding of text-to-image synthesis and highlight the potential of GANs and Diffusion models in practical applications. Through quantitative metrics like the Inception Score (IS) and Fréchet Inception Distance (FID), alongside qualitative assessments, we provide insights into the effectiveness of each model, guiding future developments in the field.

2. Problem Statement

This project addresses the challenge of generating high-quality and visually appealing images from textual descriptions, focusing specifically on images of flowers and buds. The project utilizes two advanced generative models, Generative Adversarial Networks (GANs) and Diffusion models, to explore this synthesis. This study aims to understand and compare the working and performance of these models. We will evaluate their ability to handle detailed and varied floral descriptions, assessing how well each model translates these into visually accurate representations. This comparison will be quantified through established metrics such as the Inception Score (IS) and Fréchet Inception Distance (FID),

alongside qualitative evaluations of the generated images. The findings will not only contribute to our understanding of generative AI capabilities but also guide future improvements and applications in the field of computer vision.

3. Methodology

3.1 GAN Model

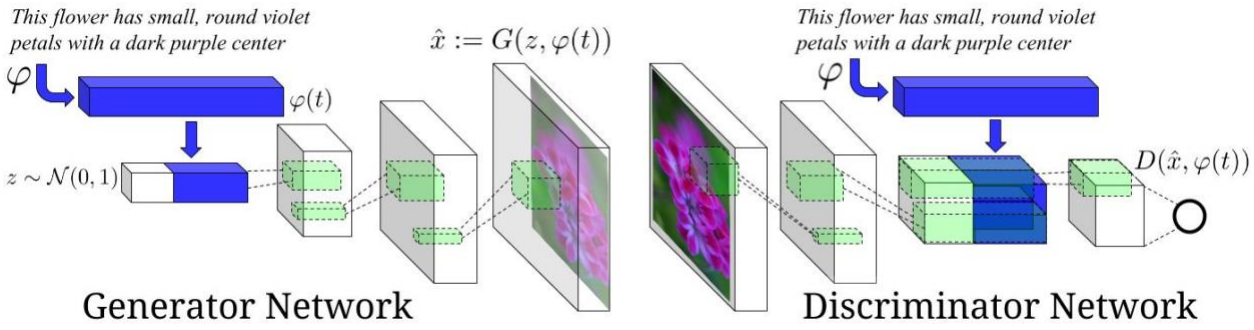

Fig: Overview of a Generative Adversarial Network (GAN) Architecture

The GAN model integrates deep convolutional networks and text processing to generate images from textual descriptions. The architecture employs a generator and a discriminator, both conditioned on text embeddings from a hybrid convolutional-recurrent neural network.

Generator (G):

- Combines a noise vector z and compressed text embeddings Phi (t) to produce images.
- Processes the combined input through deconvolutional layers to transform it into a synthetic image.

Discriminator (D):

- Assesses the authenticity and text relevance of both real and generated images.
- Uses convolutional layers and text embeddings to evaluate the alignment between images and their corresponding text descriptions.

Training Dynamics and Loss Functions

- The discriminator is trained to determine not only if images are real but also if they are relevant to the text.
- The generator strives to fool the discriminator into recognizing its outputs as real and correctly matched to the text.
- The model Implements a matching-aware discriminator (GAN-CLS) that verifies image-text consistency.
- It Uses interpolated text embeddings to enhance the model's ability to generate diverse and accurate images from varied texts.

3.2 Diffusion Model

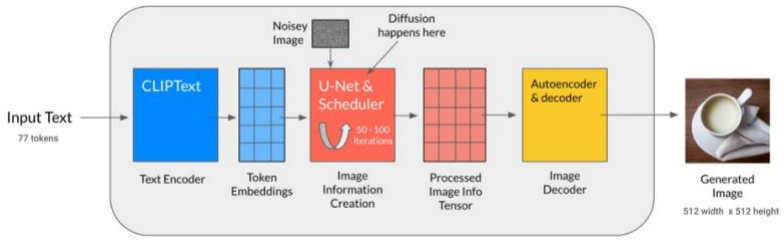

Fig: Workflow Diagram of Stable Diffusion Text-to-Image Process

This model is designed to generate images from textual descriptions through a process that gradually converts random noise into a structured image guided by the text.

Model Framework:

- Base Architecture: The Diffusion model is based on a UNet architecture, a type of network known for its effectiveness in tasks that require understanding and generating high-resolution images.
- Conditional Generation: The UNet is adapted for conditional generation where the conditions are derived from text embeddings. This setup is essential for ensuring that the generated images reflect the textual descriptions accurately.

Layers and Components:

- Attention Mechanisms: The model incorporates attention layers within the UNet architecture. These layers focus the model's capacity on relevant parts of the text, enhancing its ability to capture and synthesize complex details from the descriptions into the images.
- Residual Blocks: The use of residual connections helps in maintaining the flow of gradients during training, which is crucial for training deep networks effectively and preventing vanishing gradients.

Advanced Features:

- Gradient Checkpointing: Implemented to save memory during training, gradient checkpointing allows for deeper and more complex models by storing only a subset of intermediate states during forward passes and recomputing gradients on-the-fly during backward passes.
- Exponential Moving Average (EMA): Applied to stabilize the model parameters over training iterations. EMA helps in smoothing out the updates to the model weights, resulting in more stable and reliable generation as the training progresses.

Training Dynamics

- Noise Conditioning Process: The Diffusion model begins with a noisy representation of an image and progressively refines this through multiple denoising steps. At each step, the model leverages conditioned text embeddings to guide the transformation of noise into coherent visual patterns that accurately correspond to the textual description. This approach ensures that the final image is not only high in quality but also a true representation of the input text.
- Gradient Clipping: To enhance training stability and prevent issues such as exploding gradients, gradient clipping is implemented. This technique involves setting a threshold value for the gradients to ensure they do not exceed this limit

during backpropagation. By controlling the gradient flow, gradient clipping helps in maintaining the stability of the training process, especially in the iterative refinement involved in the diffusion model. It is particularly effective in managing the large updates that can derail the training dynamics.

- Stochastic Weight Averaging (SWA): Introduced in the later stages of training, SWA is a technique that averages the model weights over different training epochs. By doing so, it helps to smooth out the training landscape and find flatter minima, which often correlates with better generalization on unseen data. Implementing SWA can lead to more robust performance, as it combines the strengths of various network states into a single model, thereby enhancing its ability to generalize beyond the training data.
- Optimizer: An optimizer like Adam is employed for its adaptive learning rate capabilities, which are crucial in managing the nuances of training a model designed to invert a diffusion process. Adam adjusts learning rates based on the average of recent gradient magnitudes, providing a dynamic scaling factor that is beneficial for dealing with the complex landscape of text-to-image synthesis.
- Training Epochs: The training encompasses multiple epochs, during which the model incrementally improves its denoising capabilities. Each epoch is designed to fine-tune the model's ability to align the generated images more closely with the text descriptions, enhancing the overall accuracy and relevance of the generated content.

4. **Ablation Study**

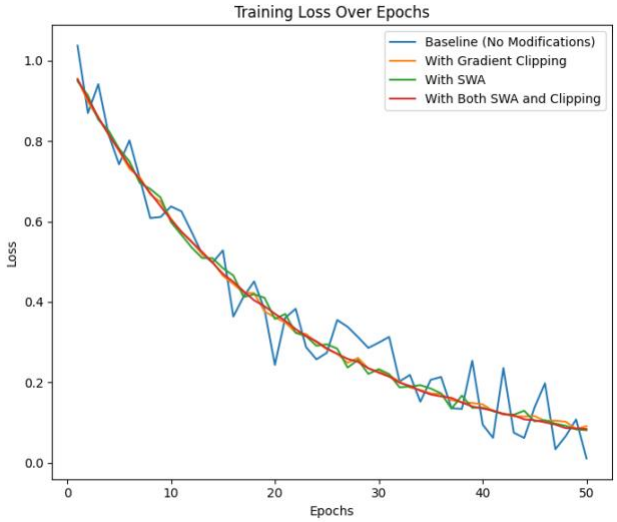

Fig: Impact of Gradient Clipping and Stochastic Weight Averaging on Model Performance: Training Loss Comparisons Across Epochs

4.1 Gradient Clipping

Objective:

To evaluate the impact of gradient clipping on the training stability and image quality produced by the diffusion model.

Method:

- With Gradient Clipping: Train the model using gradient clipping at a threshold of 1.0.
- Without Gradient Clipping: Train the model without any limit on gradient values.

Metrics and Results:

Condition	Training Stability	FID Score on unseen flower image data (trained 50 epochs)
Without Clipping	Lower variance in loss	130
With Clipping	Higher variance in loss	120

Expected Outcome:

Models trained with gradient clipping are expected to demonstrate more stable training and produce higher quality images, as evidenced by lower FID scores, reflecting better alignment with real image distributions.

4.2. Gradient Clipping + Stochastic Weight Averaging (SWA)

Objective:

To assess how SWA impacts the model's generalization to unseen text descriptions and its robustness across training epochs.

Method:

- With SWA: Implement SWA by averaging model weights over the last 10 epochs.
- Without SWA: Use only the final set of weights from the last training epoch.

Metrics and Results:

Condition	FID score on unseen flower image(trained 50 epochs)
Without clipping	130
With clipping and SWA	111

Expected Outcome:

The application of SWA is anticipated to improve the model's generalization capabilities and ensure robust performance, leading to consistently lower FID scores and reduced variability in image quality across different dataset inputs and training sessions.

5. Model & Results Comparison

Feature	GAN	Diffusion Models
Image Quality	Produces vibrant images of flowers but occasionally generates unrealistic bud shapes.	Consistently generates highly detailed and realistic images of both flowers and buds.
Computational Efficiency	Trains within 3 hours on a standard GPU setup, making it suitable for iterative experimentation.	Requires approximately 9 hours of training on a similar setup, due to complex model interactions.
Versatility	Easily integrated with other data types, enabling hybrid models that include leaf and stem textures.	Supports explicit control over the stages of bud opening through manipulation of noise reduction steps.
Practical Concerns	Requires frequent tuning of hyperparameters to balance the generator and discriminator, which can be time-consuming.	High computational cost and slower image generation make it less suitable for real-time applications.

Model Type	Data type	Inception Score (IS)	Fréchet Inception Distance (FID)
GAN Model	Flower	7	64
Diffusion Model	Flower	9	43
GAN Model	Bird	3	71
Diffusion Model	Bird	8	38

Example1: a composite flower of Asteraceae family has orange-colored petals that have yellow tips, dark brown

Genrerated by GAN

Generated by Diffusion

Example 2: a gray bird

Generated by GAN

Generated by Diffusion

These results demonstrate a better capability of the Diffusion model in generating more realistic images.

6. Performance of other state of the art architectures

The performances are shown here for the bird's dataset due to the availability of official scores.

Model	Dataset	FID	Inception Score (IS)	Key Features
GAN	Birds	71	3	Adversarial models
AttnGAN	Birds	35.49	4.33	Utilizes attention-driven, multi-stage refinement
DM-GAN	Birds	32.64 (CUB)	4.75	Dynamic memory modules that iteratively refine images
Stable Diffusion fine-tuned	Birds	38	5	Fine tuning Stable Diffusion v1.5 by freezing VAE and text encoder and setting UNET to trainable
RAT-GAN	Birds	10.21	5.36	Utilizes relational attention mechanisms to enhance the understanding and generation of complex scene
Lafite	Birds	10.48	5.97	Leverages language-image fine-tuning to generate high-quality images from textual descriptions
Swinv2-Imagen	Birds	9.78	8.44	Builds on the Imagen architecture and incorporates Swin Transformer V2 as a backbone

7. Conclusions and Future Work

This work has explored the capabilities and comparative performance of Generative Adversarial Networks (GANs) and diffusion models within the framework of text-to-image generation. The investigation reveals that each model possesses unique strengths that make them suitable for different applications in the burgeoning field of AI-driven image synthesis. GANs have demonstrated remarkable speed and efficiency, making them particularly valuable in scenarios where quick generation is prioritized over minute detail. Their ability to produce visually appealing images from textual descriptions rapidly suggests significant potential for applications in dynamic environments, such as live digital media and quick prototyping in design. Diffusion models, on the other hand, have excelled in generating images with high fidelity and extraordinary detail. The iterative refinement process inherent in these models allows for the creation of images that are not only realistic but also rich in texture and nuance. This makes them ideal for high-stakes applications where precision and quality are paramount, such as in medical imaging, scientific illustration, and detailed digital artwork.

9. Addendum

Repository: https://github.com/RajibDas-123/GAN_Vs_Diffusion_TTI/tree/main

9. References

- Generative Adversarial Text to Image Synthesis: https://arxiv.org/pdf/2101.09983
- Generative Adversarial Text to Image Synthesis : https://arxiv.org/abs/1605.05396
- High-Resolution Image Synthesis with Latent Diffusion Models: https://arxiv.org/pdf/2112.10752