Image generation with Diffusion Models

Rajdeep Roshan Sahu
230215119
r.sahu@se23.qmul.ac.uk

Abstract - This review explores image generation using efficient diffusion models in computer vision. Despite challenges like controllability and computational costs, models from OpenAI, Google AI, and Stability AI show promise. The study emphasizes the potential of diffusion models in creative expression, scientific research and various other purposes.

I. Introduction

Problem Statement: Generating creative images from human-given text prompts, using diffusion models.

Image generation is a fundamental and a creative tool provided by computer vision and Artificial Intelligence, whose applications range from creative art to scientific research. The traditional methods of image generation like, the generative adversarial network (GANs), have achieved impressive results, but they can be difficult to train and prone to generating unrealistic or unnatural images.
Thus, diffusion models have become a new method for creating images. They are more efficient to train than GANs and can generate more realistic and diverse images.

Literature: Recently, there has been a rise in the number of people that are interested in this field of study. Several research groups have developed different diffusion models, and there has been a growing body of work evaluating their performance and capabilities.

Gaps: The main challenges in this field are, understanding how diffusion models work, learning to control them effectively and to develop methods for generating images with certain characteristics and attributes.

In the further sections of the literature review, we will focus on the following aspects of diffusion models: Understanding the underlying principles of diffusion models and the different architectures that have been proposed, evaluating the performance of diffusion models using various metrics, exploring the applications of diffusion models in different fields, advantages, and disadvantages of diffusion models.

I.1 What are diffusion models?

Diffusion models are a class of neural networks that act as a powerful tool for generating images. These models are trained on huge dataset of images. Their training process involves gradually increasing the amount of noise in the images, followed by predicting the original images from the noisy versions. This teaches the model to identify the key features of an image that are robust to noise as a result. As these models are based on the idea of slowly removing noise from an image, this is done by progressively increasing the chances of generating pixels that match the image. The mathematical formula for this process is:

$$P(x) = exp(-E(x)) \quad (1)$$

where, $P(x)$ is the probability of generating an image x and $E(x)$ is the energy of image x. The energy of an image is a measure of how noisy it is. The formula for the energy of an image is:

$$E(x) = L(x) + KL(x) \quad (2)$$

where, $L(x)$ is the likelihood of the image x and $KL(x)$ is the Kullback-Leibler divergence between the image x and the prior distribution.

The likelihood of an image is a measure of how similar it is to the training data. The Kullback-Leibler divergence is a measure of how different the image x is from the prior distribution, which is a distribution that represents the model's initial guess about what the image will look like.

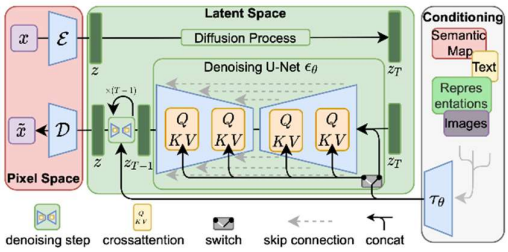

Fig.1. Workflow diagram of diffusion models

Working of the diffusion model in detail:

The diffusion model begins with a completely noisy image. Then, it gradually removes the noise from the image. In the meanwhile, it simultaneously predicts the image that was used to generate the noise. This process continues until the image is completely denoised. Key components of the diffusion model work-flow diagram:

Encoder: The encoder takes in the input image and transforms it into a latent code.

Latent space: it is used to control the image generation process. The latent space is a high-dimensional vector space that represents the possible characteristics of an image, where each dimension means a different feature of the image, like colors, shapes, and textures that can be used to create images.

Decoder: The decoder takes in the latent code and generates an image that matches the features captured by the latent code.

II. Problem Definition

Problem: Image generation using diffusion models using a neural network to generate new images. The neural network is trained on a dataset of real images.

Importance: Image generation using diffusion models is significant because it has the potential to generate new artwork, create realistic scenes for movies and video games, and generate data for scientific research.

Assumptions: Diffusion models can generate high-quality images that are not distinct from real photographs, and they can learn from and generate a wide variety of images with similar characteristics.

Objectives: To provide an overview of the diffusion models for image generation, to discuss the current and potential applications of diffusion models and to suggest directions for future research.

III. Key Works

Work 1- Diffusion Models by OpenAI for Image Generation (Dall-E 2) – OpenAI introduced diffusion models as a new approach to image generation that can generate photorealistic images.

Work 2- Imagen by Google AI – Google introduced Imagen, a diffusion model that overtook Dall-E 2 in image generation quality, photorealism and creativity.

Work 3- Stable Diffusion by StabilityAI – Stability AI launched Stable Diffusion, a diffusion model that is more stable and easier to train than previous models.

IV. Evaluation Criteria

The following evaluation criterion are commonly used to assess the performance of diffusion models:

Image Diversity: This measures the range of different images that can be generated by the model. It is evaluated using metrics Diversity Score (DS) and Variance-based Inception Score (V-IS).

Image Quality: This measures the faithfulness of the generated images to real photos. It is evaluated using metrics Inception Score (IS) and Frèchet Inception Distance (FID).

Controllability: This measures the ability of the model to generate images that match specific prompts. It is evaluated using metrics Perceptual Similarity (PS), Attribute Control Accuracy (ACA), and Text-to-Image Fidelity (TIF).

Datasets: the commonly used datasets for evaluating diffusion models are:

COCO: A large object detection and segmentation dataset.

CreativeImages: A dataset of images from the field of creative arts. (paintings, drawings, and sculptures)

ImageNet: A large dataset of natural images.

OpenImages V6: A large dataset of images with annotations for objects, scenes, and attributes.

V. Discussion

Diffusion models have shown to be a promising method for image generation when it comes to results in terms of image quality, diversity and controllability. But to reach their full potential, a few issues still need to be resolved. One main challenge is, creating techniques to produce photos with characteristics or styles, as they sometimes struggle to capture specific attributes. Also, they need to get better at producing images that stick to user specifications. Moreover, there is a need to develop more efficient training methods for diffusion models because they can be computationally expensive to train.

VI. Conclusion

Diffusion models have the power to completely transform the image creation industry, opening new possibilities for creative expression, business applications, and scientific research. They have the potential to become a vital tool for picture production in the future with more study and improvement.

References

[1] Chitwan Saharia, William Chan , Saurabh Saxena, Lala Li, Jay Whang , Emily Denton, Seyed Kamyar (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

[2] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh Pranav Shyam, Pamela Mishkin, Mark Chen (2022). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer(2022). High-Resolution Image Synthesis with Latent Diffusion Models.