Introduce to Artificial Intelligence Homework

Abstract

With the popularity of AI recently, more and more AI/CV researchers have attached a great importance to generative model-based image generation/synthesis technologies (In the following text, it will be uniformly abbreviated as "text to image", i.e., T2I). This article will focus on natural language-driven image generation/synthesis models and roughly summarize the previous related work and latest advances in this field.

Introduction

Nowadays, the technologies of image generation/synthesis are a crucial research direction in computer vision field. With the surge of artificial intelligence generated content and especially deep learning-driven generation models' boosting, the technologies of image generation/synthesis have reached an unprecedented level. These technologies can not only create original static images but also videos and 3D content. Then I will introduce the history of generation models including previous work (AE and its variants, including DAE, VAE, VQ-VAE as well as diffusion model series) and latest advances (OpenAI DALL · E series and Google Imagen series).

1 Previous Related Work

1.1 GAN

This work was proposed by Ian Goodfellow et al. in 2014, with its core innovation being in the firstly introduction of the "adversarial learning" mechanism. (I have read Ian's book on deep learning, so he left a deep impression on me). The **GAN** model has two main parts: one is the *Generator* and the other is *Discriminator*. The objective of *Generator* is to learn the distribution of real data and generate fake data that are as realistic as possible (for example, image) while the objective of *Discriminator* is to determine whether the images are real data or fake data generated by *Generator*. The ultimate goal of the **GAN** model is to train these two parts so that the *Generator* has the ability to generate fake images that cannot be distinguished by *Discriminator*.

Advantages

- High-quality data generation

- No explicit probability modeling required

Disadvantages

- Training instability
- High data requirement
- Lack of image diversity

1.1.1 Wasserstein GAN (WGAN)

WGAN addresses the training instability of vanilla GANs by introducing the Wasserstein distance (Earth Mover's Distance) as a new loss metric. This approach provides smoother gradients and alleviates mode collapse. WGAN replaces the discriminator with a "critic" that scores the realness of samples, and enforces a Lipschitz constraint, typically via weight clipping or gradient penalty (WGAN-GP). As a result, WGANs are more stable and easier to train, especially on complex datasets.

1.1.2 Conditional GAN (CGAN)

CGAN extends the GAN framework by conditioning both the generator and discriminator on auxiliary information, such as class labels or attributes. This allows the model to generate images corresponding to specific categories or features, enabling controlled image synthesis. The conditioning is usually implemented by concatenating the label information with the input noise vector and/or the input to the discriminator.

1.1.3 BigGAN

BigGAN, proposed by DeepMind in 2018, scales up the GAN architecture to achieve state-of-the-art image synthesis on large and complex datasets like ImageNet. BigGAN uses larger batch sizes, deeper networks, and advanced regularization techniques. It also incorporates class-conditional information and spectral normalization to stabilize training. BigGAN demonstrates that scaling up GANs leads to significant improvements in image fidelity and diversity, but also requires careful tuning and substantial computational resources.

1.2 AE Series

Due to the information density of images is low, there is a significant pixel redundancy. For example, the concept of "a cat" can be expressed in natural language with just a few words, whereas representing it as an RGB image requires hundreds or thousands of pixels. And a work named **MAE** proposed by Kaiming He et al. in 2021 has proved that if an image is masked by 75% region, it can still be roughly reconstructed by a transformer-based image model. This further demonstrates that images are a low-information density modality.

Under this prior condition, AE series' main algorithm is to generate an image from a latent vector, whose dimensionality is much lower than the original image.

1.2.1 AutoEncoder (AE)

The inputs for **AE** are the original images. Each image is firstly encoded into a latent vector by an encoder, and then decoded into a reconstructed image by a decoder. The loss function used is mean squared error (MSE) Loss.

1.2.2 Denoising AutoEncoder (DAE)

The main difference between an **AutoEncoder (AE)** and a **Denoising AutoEncoder (DAE)** is that, in DAE, noise is deliberately added to the input images during training, and the model is trained to reconstruct the original, noise-free images from these corrupted inputs. Hence the name, **DAE**.

Though **AE** and **DAE** are good at reconstructing the original images, reconstruction is the only thing they can do. The latent vector, also called bottleneck, is not modeled as a probability distribution. Therefore, it cannot be randomly sampled to generate new images.

1.2.3 Variational AutoEncoder (VAE)

To solve the disadvantages of **AE** and **DAE**, **Variational AutoEncoder** not only compresses the input image into a latent vector, but also requires that these latent vectors follow a probability distribution (usually a standard normal distribution), while retaining the well-designed encoder-decoder architecture of **AE** and **DAE**.

It should be noted that the loss of a VAE contains not only MSE loss but also KL loss. Specifically, the KL loss is calculated as:

$$\text{KL\_loss} = -0.5 \times \sum \left(1 + \log(\sigma^2) - \mu^2 - \sigma^2\right) \tag{1}$$

Note that μ and σ^2 represent the mean and variance, respectively.

VAE has many advantages, such as being able to generate new data samples rather than just reconstructing the input, and providing a clear probabilistic interpretation. However, it also has some drawbacks: on the other hand, since **VAE** models the latent space using continuous probability distributions, it cannot effectively handle discrete data. At the same time, when dealing with more complex or large-scale data, the scalability of the model can be limited, and the quality of generated samples is sometimes inferior to that of other generative models.

1.2.4 Vector Quantized Variational AutoEncoder (VQ-VAE)

VQ-VAE uses a codebook to discretize the latent space, instead of modeling it with a normal distribution as in a standard **VAE**. This approach allows the features produced by the encoder to be quantized into discrete representations. While this enables effective compression and discrete modeling, **VQ-VAE** alone cannot directly generate new images. To sample or generate new images, an additional prior network (such as PixelCNN or Transformer) is required to model the distribution over the discrete latent variables. Without such a prior, **VQ-VAE** can only be used for tasks like reconstruction or feature extraction, rather than true generative modeling.

It should be noted that the loss function of a **VQ-VAE** consists of three parts: the reconstruction loss (MSELoss), the codebook loss, and the commitment loss.

$$\text{VQ-VAE\_loss} = \text{recon\_loss} + \text{vq\_loss} + \beta \times \text{commit\_loss} \tag{2}$$

For its **Prior** model, the loss function is CrossEntropyLoss, since it is an autoregressive model.

$$\text{Prior\_loss} = \text{CrossEntropyLoss} \tag{3}$$

1.3 Diffusion Series

1.3.1 Diffusion

The original diffusion models are defined as follows:

Forward diffusion: Given an image, a small amount of noise is added at each step, for a total of T steps. If T is very large, the image will eventually become pure noise, specifically isotropic Gaussian noise.

Reverse diffusion: Starting from random noise, a neural network is used to gradually remove the noise, step by step, until an image is reconstructed.

Because T is typically quite large (e.g., T = 1000 in the original diffusion models), both training and inference of diffusion models are much more time-consuming than models like GANs, as multiple forward passes are required.

To maintain the same input and output image size during the denoising process, the original diffusion model adopts a shared-parameter U-Net, which is an encoder-decoder structure. To further improve image reconstruction quality, skip connections and attention mechanisms were introduced into the U-Net.

Although the concept of diffusion models was proposed over a decade ago, it wasn't until 2020, with the introduction of DDPM, that diffusion models began to gain widespread attention.

1.3.2 DDPM

DDPM (Denoising Diffusion Probabilistic Models) introduced two main innovations:

1. **Noise prediction:** In the denoising process, instead of directly predicting the clean image at each step, DDPM predicts the added noise (analogous to the residual in ResNet). Additionally, time embeddings are incorporated into the U-Net, which helps the model generate and sample images more effectively. This works because the U-Net uses shared parameters, and time-step embeddings enable the model to produce step-specific outputs. For example, in early denoising steps, the model is expected to generate only the coarse outline of objects, while finer details and high-frequency features are gradually restored as the steps progress.

2. **Variance simplification:** While predicting a Gaussian distribution typically requires modeling both the mean and variance, DDPM shows that it suffices to predict only the mean, with the variance set as a constant. This simplification not only maintains strong performance but also eases model optimization.

1.3.3 Improved DDPM

OpenAI made several improvements to DDPM, with three key contributions:

1. **Learnable variance:** While the original DDPM used a fixed variance, the OpenAI team made the variance a learnable parameter, which improved the quality of generated and sampled images.

2. **Noise schedule optimization:** The noise addition schedule was changed from a linear schedule to a cosine schedule (similar to how learning rate schedules work), resulting in significantly better performance.

3. **Scalability and classifier guidance:** OpenAI demonstrated that DDPM scales very well with large and complex models. In the "Diffusion Beats GAN" paper (**DALL-E 2**'s second and third authors), they increased the model's depth and width, and added more self-attention heads. Since single-scale attention was insufficient, they introduced multi-scale attention, making the models both larger and more complex. Furthermore, they proposed a classifier guidance method, which reduced the number of sampling steps to around 25. These innovations laid the foundation for subsequent research in this field.

Since then, diffusion models have risen to prominence and become one of the hottest research directions in image generation.

1.3.4 Classifier Guidance

1. Let's focus on the transition from X_t to X_{t-1}. If we add a large amount of noise to images from ImageNet and use these noisy images to train a classifier, the resulting classifier can facilitate the denoising process for the diffusion model. More generally, before each denoising step, we can use the classifier to make a prediction, obtain its gradient with respect to the input, and use this gradient to guide the diffusion process in the correct direction during parameter updates.

2. Following the success of classifier guidance, some researchers began to use a CLIP model as the guidance signal instead of a traditional classifier. This approach proved to be very effective, allowing both text and images to serve as guidance signals. Different guidance signals lead to different types of generation tasks, including text-driven image generation, image reconstruction, and image style transfer.

1.3.5 Classifier-free Guidance

If we assume the guidance signal is text, then during training, the model needs to perform forward passes both with and without the conditioning signal. In this way, the model learns the impact that the text signal has on the output. Although this method is computationally expensive, its effectiveness cannot be ignored. After being proposed, this approach was adopted in **GLIDE**, **DALL-E 2** and **Imagen**.

1.3.6 GLIDE

GLIDE incorporates techniques such as DDPM, improved DDPM, and classifier-free guidance, and is a diffusion model with 3.5 billion parameters. It later evolved into DALL · E 2.

1.4 DALL-E series

Originally, **VQ-VAE** used **PixelCNN** as its prior for sampling or generating new images, taking advantage of **PixelCNN**'s autoregressive nature. However, OpenAI later replaced **PixelCNN** with their signature autoregressive model, **GPT**, to further enhance the quality and flexibility of image generation.

1.4.1 DALL-E

The **DALL-E** model, developed by OpenAI, is a text-to-image generation model. During training, the text prompt is first encoded into text tokens, and the image is encoded into image tokens using a **VQ-VAE** encoder. These two sequences of tokens are concatenated and fed into a **GPT** model, which learns to perform autoregressive modeling over the combined sequence. During inference, the process is simpler: the text prompt is encoded into text tokens, which are then passed to the **GPT** model to autoregressively predict the corresponding image tokens. Finally, these image tokens are decoded by the **VQ-VAE** decoder to generate a pixel-level image.

1.4.2 DALL-E 2

The training datasets of **DALL-E 2** consist of pairs (x, y) of images and their corresponding captions y. Given an image x, let z_i and z_t be its CLIP image and text embeddings, respectively.

$$P(z_t|y) = \text{Prior}(y) \quad \text{and} \quad P(x|z_t, y) = \text{Decoder}(z_t, y) \tag{4}$$

The generative stack uses two components to produce images from captions: a **prior** model and a **decoder** model. The prior produces CLIP image embeddings z_i conditioned on caption y. The decoder produces images x conditioned on CLIP image embeddings z_i and text caption y.

$$P(x|y) = P(x, z_i|y) = P(x|z_i, y)P(z_i|y). \tag{5}$$

The DALL-E 2 model uses CLIP guidance and classifier-free guidance to improve the quality of generated images but with a high computational cost. To generate high resolution images, DALL-E 2 uses two diffusion upsample models. There 2 choices for prior model: a diffusion prior model ans a autoregressive prior model. For DALL-E 2, directly predicting the unnoised z_i is better than predicting the noise ϵ. Maybe the reason is that the target z_i is more suitable for the DALL-E 2 model's architecture.

1.5 Imagen

Imagen is a text-to-image diffusion model proposed by Google Research, which achieves state-of-the-art performance on several image generation benchmarks. The core idea of Imagen is to leverage large

pre-trained language models for text encoding and combine them with a powerful diffusion-based image decoder.

Unlike DALL-E 2, which uses CLIP for both text and image embeddings, Imagen utilizes a large frozen T5 language model to encode the text prompt, arguing that high-quality language understanding is crucial for text-to-image generation. The text embedding produced by T5 is then fed into a series of diffusion models to generate images.

The overall pipeline of Imagen consists of three main components:

1. **Text Encoder:** A large pre-trained T5 model is used to encode the input text prompt into a dense embedding.

2. **Base Diffusion Model:** The text embedding is concatenated with noise and passed into a U-Net-based diffusion model to generate a low-resolution image (e.g., 64×64).

3. **Super-Resolution Diffusion Models:** Two cascaded diffusion models are used to progressively upsample the image to higher resolutions (e.g., 256×256 and 1024×1024), each conditioned on the text embedding and the previous lower-resolution image.

Imagen introduces several key innovations: Firstly, Instead of using cross-attention, Imagen injects the text embedding into the diffusion U-Net via a simple concatenation and FiLM (Feature-wise Linear Modulation) layers, which improves both efficiency and performance. Secondly, similar to GLIDE and DALL-E 2, Imagen adopts classifier-free guidance to further enhance the fidelity and relevance of generated images.

Experimental results show that Imagen achieves unprecedented photorealism and language-image alignment, outperforming previous models such as DALL-E 2 and GLIDE on various benchmarks.

1.6 Summary of previous work

These work are all proposed before 2022.10 when the chatgpt firstly released. But they totally proved:

1. **Classifier-free guidance** is the answer. It is a very useful technique to improve the performance of generative models.

2. **Diffusion model** is the answer. It is a very powerful generative model, which can be used to high-quality images with both variability and fidelity.

3. **Scaling** is the answer. Whatever the architecture or the training strategy or the loss function, the generative models are all need to "Scale up" to get the better performance.

2 Latest Advances

Since late 2022, the field of text-to-image generation has witnessed rapid and transformative advances, driven by both academic research and industry applications. Below, we summarize several of the most influential developments:

2.1 Stable Diffusion and the Open-Source Revolution

Stable Diffusion, released by Stability AI in August 2022, marked a turning point by making high-quality text-to-image diffusion models accessible to the public. Unlike previous proprietary models, Stable Diffusion is fully open-source, allowing researchers, developers, and artists to fine-tune, deploy, and build upon the model freely. Its lightweight architecture enables efficient inference on consumer GPUs, greatly lowering the barrier to entry. The open-source ecosystem around Stable Diffusion has fostered rapid innovation, including community-driven improvements, plug-ins, and creative applications.

2.2 Midjourney and Community-Driven Innovation

Midjourney is a commercial text-to-image platform that has gained popularity for its unique artistic style and user-friendly interface. By leveraging Discord as its primary interaction channel, Midjourney has built a vibrant community where users can collaboratively explore prompt engineering and share results. The model is known for its stylized, imaginative outputs, and frequent updates based on user feedback.

2.3 DALL · E 3 and Prompt Engineering

OpenAI's DALL · E 3, released in 2023, further improved the fidelity, coherence, and controllability of generated images. DALL · E 3 is notable for its deep integration with large language models (LLMs), enabling users to describe complex scenes in natural language and receive highly relevant images. The model demonstrates strong prompt following and can handle nuanced instructions, making prompt engineering a critical skill for users.

2.4 Imagen Editor and Controllable Generation

Google's Imagen Editor extends the Imagen family by introducing fine-grained, interactive image editing capabilities. Users can specify edits using natural language, such as modifying objects, styles, or backgrounds within an image. This controllability is achieved by conditioning the diffusion process on both the original image and the edit instructions, opening new possibilities for creative workflows and content customization.

2.5 Multimodal Large Models: GPT-4V, Gemini, and Beyond

The latest trend is the emergence of multimodal large models, such as OpenAI's GPT-4V and Google's Gemini, which can process and generate both text and images (and even audio or video). These models unify language and vision understanding, enabling tasks like visual question answering, image captioning, and cross-modal reasoning. Their generalization ability and scalability are pushing the boundaries of what generative AI can achieve.

2.6 Future Trends and Challenges

Despite remarkable progress, several challenges remain: improving controllability, reducing biases, ensuring safety, and scaling to higher resolutions and more modalities. The integration of generative models with real-world applications (e.g., design, education, entertainment) is expected to accelerate, while ethical and societal considerations will become increasingly important.