확산 모델 기반 텍스트 정보를 이용한 이미지 생성 모델 연구

김소정, 문아성, 이재성
중앙대학교 AI학과
e-mail: *sojeong6894, kove1230, curseor@cau.ac.kr*

A study of text guided image generation based on diffusion model

Sojeong Kim, A Seong Moon, Jaesung Lee
Department of Artificial Intelligence
Chung-Ang University

Abstract

This paper discusses the use of diffusion models in text to image generation tasks. Diffusion models are considered promising generative models that can generate photorealistic images from text descriptions. Introducing the background of diffusion models with classifier guidance and implemented space, we will show the results of inference experiments of three recent diffusion models: GLIDE, VQ-diffusion, and Stable Diffusion. With these results, we will discuss the future research directions of text to image generation using diffusion model.

I. Introduction

Text to image generation is the process of generating realistic images from textual descriptions using neural networks. Recently, diffusion models have emerged as a promising generative modeling framework, pushing the state of the art on text to image generation tasks. It is possible to adjust the level of detail and style of the generated images which is useful for generating a variety of realistic images from text descriptions. In this paper, we discuss a diffusion model that handles the text to image generation tasks such as GLIDE [7], Imagen [8], VQ-diffusion [9], Stable diffusion [10]. This paper is organized as follows. Section 2 introduces the background of diffusion model including the classifier guidance type, which is important for the text to image generation task. And then, the four diffusion models mentioned above are divided by the criteria of implementation space and guidance type, and each model is briefly described. Section 3 presents the results of inference experiments with the MS-COCO[11] 2014 dataset for three diffusion models: GLIDE, VQ-diffusion, and Stable Diffusion. As evaluation metrics, five quantitative indicators: IS[12], FID[13], Clip score[14], PSNR[15], and SSIM[16] are used. Finally, Section 4 discusses the results of inference experiments and the future research directions using diffusion model in text to image generation tasks.

II. Related Works

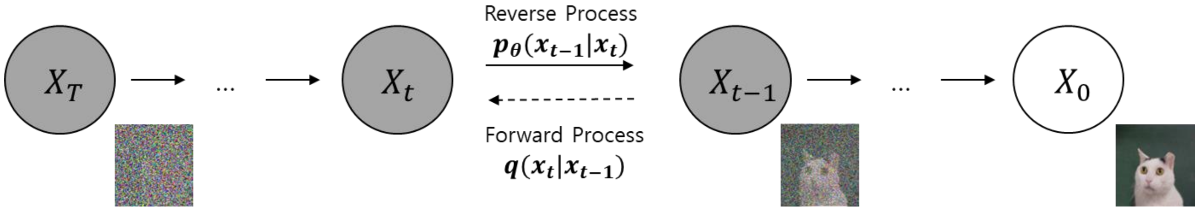

Figure 1. The graphical model of Forward (q) and backward (p) diffusion process

2.1 Background

Diffusion model used in deep generative modeling was first introduced in 2015[1]. This paper proposed the use of a Markov chain which implements the forward diffusion process by adding Gaussian noise to the data, and a diffusion probabilistic model (DPM) to implement the backward diffusion process, which denoises the data in many small denoising steps. The concepts of the DPM has been extended to several variants such as Noise Conditional Score Networks (NCSN)[2] and Denoising Diffusion Probabilistic Models (DDPM)[3].

In Noise Conditional Score Networks (NCSN)[2], a score-based generative modeling method that generates samples using gradients of the data distribution estimated by score matching through Langevin dynamics was proposed in 2019[2]. In the assumption that the sample x follows a Gaussian distribution, the score network s_θ expressed as an expression for Gaussian noise as Eq.(1).

$$s_\theta(x_t,t) \approx \nabla_{x_t} log\, q(x_t) = -\frac{\epsilon_\theta(x_t,t)}{\sqrt{1-\bar{\alpha}_t}} \qquad (1)$$

NCSN sampling using the score network $s\theta$ is shown in Algorithm 1.

```
Algorithm 1. NCSN Sampling
Require: ${\sigma_i}_{i=1}^{L}, \epsilon, T$
Initialize $x_0$
for $i = 1, \ldots, L$ do
    $\alpha_i = \epsilon \cdot \sigma_i^2/\sigma_L^2$
    for t= $1, \ldots, T$ do
        $z_t \sim \mathcal{N}(0, I)$
        $x_t = x_{t-1} + \frac{\alpha_i}{2} s_\theta(x_{t-1}, \sigma_i) + \sqrt{\alpha_i} z_t$
    end for
    $x_0 = x_T$
end for
return $x_T$
```

In case of Denoising Diffusion Probabilistic Model (DDPM) proposed in 2020[3], without directly learning the score network s_θ, a form similar to denoising score matching[4] can be obtained by inducing the objective function of the existing diffusion model[1], and finally the following sampling Algorithm 2 is obtained.

2.2. Conditional Diffusion Model

2.2.1 Classifier Guided Diffusion

For conditional diffusion process guided by class information y, a gradient of a classifier $\nabla_x \log f_\phi(y|x_t)$ conditioned on noisy image x_t was introduced in 2021[5]. By modifying the noise prediction at each step, the diffusion model is guided according to the desired conditioning class information y. Based on NCSN, DDPM with Eq.(1), and other several formulas included in the paper, it leads to a new classifier guided predictor $\bar{\epsilon}_\theta(x_t,t)$ with a classifier guidance w to add a weight to the delta part, as shown in Eq.(2). For more detail, the following sampling Algorithm 3 is obtained.

$$\bar{\epsilon}_\theta(x_t,t) = \epsilon_\theta(x_t,t) - \sqrt{1-\bar{\alpha}_t} w \nabla_{x_t} \log f_\phi(y|x_t) \qquad (2)$$

2.2.2 Classifier free Guidance

Conditional diffusion can be achieved without an independent classifier by training a conditional diffusion model together with an unconditional model in 2021[6]. According to the paper, parameterizing denoising diffusion model p to score estimator ϵ_θ, with a classifier guidance w, classifier free guidance $\bar{\epsilon}_\theta(x_t,t,y)$ can be written as Eq.(3).

$$\bar{\epsilon}_\theta(x_t,t,y) = \epsilon_\theta(x_t,t,y) - \sqrt{1-\bar{\alpha}_t} w \nabla_{x_t} \log p(y|x_t)$$

$$= (w+1)\epsilon_\theta(x_t,t,y) - w\epsilon_\theta(x_t,t) \qquad (3)$$

Algorithm 2. DDPM Sampling

```
$x_T \sim \mathcal{N}(0, I)$
for t= $T$,...,1 do
    $z \sim \mathcal{N}(0, I)\ if\ t > 1, else\ z = 0$
    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$
end for
return $x_0$
```

2.3 Representative Diffusion Models

2.3.1 Diffusion on pixel space

GLIDE[7], the guided diffusion model, used both CLIP guidance and classifier free guidance. For classifier guidance method, they used CLIP guidance replacing the classifier with a CLIP model trained on noised images and for classifier free guidance, they replaced original class label with text. In the Experimental section, they found that human evaluators preferred images generated from classifier free guidance over CLIP guidance in terms of the sample's photorealism and caption similarity. The model architecture consists of a U-Net architecture with residual blocks and attention blocks, and the text encoder is set to a transformer with 24 residual blocks and a width of 2048, resulting in roughly 1.2 billion parameters.

Imagen[8] is a pixel space diffusion model generating images directly from the high-dimensional pixel level adopting classifier free guidance. The Imagen model has a text encoder that can be pretrained on either image-caption pairs such as CLIP or text-only corpora like GPT, BERT, T5. In Experimental results, they found that increasing the size of the language model, such as T5, is more effective for improving both image fidelity and image-text alignment than increasing the diffusion model, UNet size. It works because text-only corpus is larger than paired image-text data, the large language models can be exposed to a diverse and extensive range of text data.

2.3.2 Diffusion on latent space

VQ-Diffusion(Vector Quantized Diffusion) [9] is a conditional latent diffusion model incorporating classifier guidance for better control over the generated images. It combines the Denoising Diffusion Probabilistic Model(DDPM) [1, 3] and the pre-trained Vector Quantized Variational Autoencoder (VQ-VAE), eliminating the unidirectional bias and avoiding accumulated prediction errors. VQ-Diffusion's diffusion process is implemented in a quantized latent space, and they found that the latent space model is suitable for text to image generation task. The experimental results indicate that the VQ-Diffusion outperforms conventional autoregressive (AR) models with similar parameter sizes.

Algorithm 3. Classifier guided diffusion sampling, given a diffusion model $(\boldsymbol{\mu_\theta(x_t)}, \boldsymbol{\Sigma_\theta(x_t)})$, classifier $\boldsymbol{p_\phi(y|x_t)}$, and gradient scale s.

```
Input: class label y, gradient scale s
$x_T \sim$ sample from $\mathcal{N}(0, I)$
for t= $T$,...,1 do
    $\mu, \Sigma \sim \mu_\theta(x_t), \Sigma_\theta(x_t)$
    $x_{t-1} \sim$ sample from $\mathcal{N}(\mu + s\Sigma\nabla_{x_t}\log p_\phi(y|x_t), \Sigma)$
end for
return $x_0$
```

Stable Diffusion[10] is a classifier free guidance diffusion model that operates in the pixel space. It compresses images using a variational autoencoder (VAE) into a lower dimensional latent embedding, and then by using VAE decoder, generates latent representations of images from the latent embeddings. It uses CLIP's pre-trained text encoder inspired by Imagen to condition the model on text prompts. The results of experiments demonstrate that implementing diffusion process on the latent space provides better results in reducing complexity and preserving image fidelity compared to processing directly on the high-dimensional pixel space.

Ⅲ. Experiments

3.1 Implements

As shown in Table 3, the following three models: GLIDE, VQ-diffusion and Stable Diffusion were selected according to the implemented space and guidance type. In case of Imagen, an official pre-trained model was not available, so it was replaced with a classifier-free version of GLIDE.

Table 1. Classification of Diffusison Models by implemented space and guidance type

Model	Pixel based		Latent based		Scale
	CG	CFG	CG	CFG	
GLIDE	✔	✔			3.0
VQ-diffusion			✔		✘
Stable Diffusion				✔	7.5

CG; Classifier Guidance, CFG; Classifier-free guidance

3.2 Dataset

MS COCO 2014 Dataset[11] contains a large number of diverse images with multiple objects and scenes, and has been used in experiments of many text to image generation models, so we chose MS COCO 2014 dataset for inference experiments of diffusion models. In addition, since there was no caption data available for the images in the test set, we used validation set. The MS COCO 2014 dataset's validation set includes 40,504 images, and each image has 5 captions. Therefore, we randomly selected one from multiple captions for each image.

3.3 Evaluation metrics

To evaluate different models in terms of generated images quality and diversity, we selected 5 methods, Inception Score (IS)[12], Frechet Inception Distance (FID)[13], CLIP score[14] for text relevance and Peak Signal-to-Noise Ratio (PSNR)[15], Structural Similarity Index (SSIM)[16] for high quality of the generated images. As shown in Table 2, the evaluation results were calculated from 10,000 randomly selected generated images from each model.

3.4 Experimental Results

The evaluation results of inference experiments on MS-COCO dataset, randomly extracted 10,000 generated images are shown in Table 2. Larger IS, PSNR, SSIM and smaller FID mean higher image fidelity and CLIP scores mean text-image alignment that is important for text to image generation tasks. As a result of the experiment, Stable Diffusion scored the best in IS, FID, SSIM, and GLIDE(classifier-free guidance) in PSNR. Therefore, using the Stable Diffusion, which is a latent space-based diffusion model, seems to be more effective and efficient in terms of increasing the fidelity of the image and reducing the complexity of model operation.

Table 2. Results of MS COCO

MS COCO					
Model	IS↑	FID↓	CLIP score↑	PSNR↑	SSIM↑
GLIDE (CG)	24.64	44.18	17.50	8.37	**0.13**
GLIDE (CFG)	14.21	98.21	17.29	**8.66**	0.04
VQ-diffusion	24.29	37.04	29.46	8.41	0.12
Stable diffusion	**31.74**	**19.83**	**29.56**	8.54	**0.13**

10,000 randomly selected images generated from each diffusion model were used to calculate each metric.

Since this inference experiment compared images generated from randomly selected text among multiple captions and the metrics were calculated from 10,000 generated images, the results have a significant difference from the performance of each officially announced model. However, it is meaningful in that several promising diffusion models, such as GLIDE, VQ-diffusion, and Stable diffusion, were compared under a fairly generalized conditions with some limitations, rather than in an ideal environment.

Ⅳ. Conclusions

This paper introduces several promising diffusion models for generating images from text descriptions. Providing an overview of the diffusion models in terms of classifier guidance type and implemented space, inference experiments on three promising diffusion models (GLIDE, VQ-diffusion, and Stable Diffusion) were performed using the MS COCO 2014 dataset. The quality of the generated images and the degree of text relevance were measured through 5 quantitative evaluation metrics such as IS, FID, CLIP score, PSNR, and SSIM. As a result of the inference experiment, the performance of Stable Diffusion was superior to other models in terms of image fidelity and text relevance, so it could be estimated that it has excellent efficiency and performance in the text to image generation task. Therefore, in future research on text−image generation using diffusion models, we would like to do more research on the latent based diffusion models in more depth such as Stable Diffusion, improving the quality of the generated image and reducing the computational complexity which is a chronic problem of diffusion model.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021−0−01341−003,Artificial Intelligence Graduate School Program(Chung−Ang University))

References

[1] Jascha Sohl−Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In ICML, 2015.

[2] Yang Song, Stefano Ermon. Generative modeling by estimating gradients of the data distribution. arXiv:1907.05600, 2020b.

[3] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In ICLR. OpenReview.net, 2021.

[4] Pascal Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7):1661–1674, 2011.

[5] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. CoRR, abs/2105.05233, 2021.

[6] Jonathan Ho and Tim Salimans. Classifier−free diffusion guidance. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.

[7] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text−guided diffusion models. CoRR, abs/2112.10741, 2021.

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes et al., "Photorealistic text−to−image diffusion models with deep language understanding," arXiv preprint arXiv:2205.11487, 2022.

[9] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text−to−image synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 696–10 706.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High−resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 684–10 695.

[11] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L. Microsoft coco: Common objects in context. In: ECCV(2014)

[12] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen,X. Improved techniques for training gans. NeurIPS. 2016

[13] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. Gans trained by a two time−scale update rule converge to a local nash equilibrium. NeurIPS. 2017

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference−free Evaluation Metric for Image Captioning. arXiv preprint arXiv:2104.08718, 2021.

[15] Fernando A. Fardo, Victor H. Conforto, Francisco C. de Oliveira, Paulo S. Rodrigues. A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms. arXiv preprint arXiv:1605.07116, 2016

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600−612, 2004.