Text to Image Generation using Stable Diffusion

Shubham Vatsal, Rajat Narlawar, and Yulu Qin

Courant Institute of Mathematical Science, New York University, New York, USA
{sv2128,rsn9148,yq810}@nyu.edu

Abstract. Text-to-image generation is a challenging problem in the field of artificial intelligence, as it involves generating visual data from linguistic input. In this paper, we investigate the use of the latest stable diffusion model for text-to-image generation. Stable diffusion models are a variant of diffusion models that have been shown to be more effective at generating high-quality images from text descriptions. In this work, we fine-tune a stable diffusion model on CIFAR-10 and Oxford-102 datasets and generate images corresponding to their mapped textual inputs. Once the model has been fine-tuned, we evaluate its performance on a downstream classification task for CIFAR-10 and manually analyze as well as calculate Frechet Inception Distance (FID) for both Oxford-102 and CIFAR-10. The discussed work apart from providing insight into the capabilities of the used stable diffusion model for text-to-image generation also publishes base pipeline code which can be easily used to fine-tune any other dataset.

Keywords: Stable Diffusion · Text-to-Image Generation · Generative Model · Classification · CIFAR-10 · Oxford-102 ·

1 Introduction

It is practically impossible to convince an astronaut to ride a horse in his or her space suit and then take their picture or it is impossible to make one's favourite basketball player ride a dinosaur. Is there a way we can visually perceive these imaginary pictures. Of course there have been many video editing tools from ages but it requires a lot of effort even using those tools to create such an unreal image. It is challenging to create such fictitious scenes, which requires synthesizing examples of specific subjects in novel settings such that they organically and flawlessly fit into the scene. But large deep learning based generative models successfully handle this challenge.

The improvement of quality of generated images rendered by various generative models over the years can be seen in Fig 1.

Text-to-image generation can be defined as construction of images according to natural language descriptions in an automated environment. Descriptive language phrases are an intuitive and flexible way to describe visual notions for creating images when compared to other types of input such as sketches, audio and object masks. Learning from unstructured description and managing

Fig. 1. Improvement of Quality of Images Generated Over the Years

the differing statistical features between vision and language inputs are the primary challenges for text-to-image synthesis. It is a fundamental problem in many applications, such as art generation and multimedia content creation. Text-to-image models offer unprecedented freedom to guide creation through natural language. It has also made multimodal domain across vision and language as one of the most actively researched areas [18, 17, 36, 25, 5, 1, 28, 33, 34].

The rest of the paper is organised in the following way. Section 2 talks about the related work. We illustrate the working of the used stable diffusion model in section 3. Section 4 concentrates on the experiments we conducted and the corresponding results we achieved. The final section gives a summary of this paper.

2 Related Work

There have been many work done in the field of image generation conditioned on various factors using different kinds of generative models. [35] talks about the image-to-image generation by modelling a distribution of possible outputs in a conditional generative modeling setting. They propose that the ambiguity of the mapping is distilled in a low-dimensional latent vector, which can be randomly sampled at test time. [27] investigates generation of images from visual attributes. The authors model an image as a composite of foreground and background and develop a layered generative model with disentangled latent variables that can be learned end-to-end using a variational auto-encoder. [24] generates emojis from images using their proposed Domain Transfer Network which employs a compound loss function that includes a multi-class Generative Adversarial Network (GAN) loss, an f-constancy component, and a regularizing component. [4] uses conditional generative adversarial networks to generate images from audio. [12] introduces a word-level spatial and channel-wise attention-driven generator that can disentangle different visual attributes and allow the model to focus on generating and manipulating sub-regions corresponding to the most relevant words. They further discuss a word-level discriminator to provide fine-grained supervisory feedback by correlating words with image regions, facil-

itating training an effective generator which is able to manipulate specific visual attributes without affecting the generation of other content.

A few years back, GANs [6] appeared as one of the most promising approaches for generative modeling. [3] presents variational generative adversarial networks, a general learning framework that combines a variational auto-encoder with a generative adversarial network, for synthesizing images in fine-grained categories, such as faces of a specific person or objects in a category. [30] proposes double attention which simultaneously leverages the context of the local and the shifted windows, leading to improved generation quality. Lifelong GAN [29] employs knowledge distillation to transfer learned knowledge from previous networks to the new network. This makes it possible for them to perform image-conditioned generation tasks in a lifelong learning setting.[14] talks about a novel framework Semantic-Spatial Aware GAN for synthesizing images from input text. The authors introduce a simple and effective Semantic-Spatial Aware block which learns semantic-adaptive transformation conditioned on text to effectively fuse text features and image features and learns a semantic mask in a weakly-supervised way that depends on the current text-image fusion process in order to guide the transformation spatially.

Diffusion models have become quite popular nowadays for beating GANs at the task of generative modelling. [9] proposes cascaded diffusion model comprising of a pipeline of multiple diffusion models that generate images of increasing resolution, beginning with a standard diffusion model at the lowest resolution, followed by one or more super-resolution diffusion models that successively upsample the image and add higher resolution details. [21] discusses their simple implementation of image-to-image diffusion model and claim it outperforms strong GAN and regression baselines on all tasks, without task-specific hyperparameter tuning, architecture customization, or any auxiliary loss or sophisticated new techniques needed. The authors further uncover the impact of an L2 vs. L1 loss in the denoising diffusion objective on sample diversity, and demonstrate the importance of self-attention in the neural architecture through empirical studies. [7] presents the vector quantized diffusion (VQ-Diffusion) model for text-to-image generation based on a vector quantized variational autoencoder (VQ-VAE) whose latent space is modeled by a conditional variant of Denoising Diffusion Probabilistic Model (DDPM). [20] leverages the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss which enables synthesizing the subject in diverse scenes, poses, views, and lighting conditions that do not appear in the reference images. [19] achieves state-of-the-art synthesis results on image data. It introduces cross-attention layers into the model architecture which allows diffusion models to become flexible generators for general conditioning inputs such as text or bounding boxes allowing high resolution synthesis in a convolutional manner.

There have been several prior work published on text-to-image generation using GANs and diffusion models. GANs came up with promising results on text-to-image generation [31, 32]. AttnGAN [26] proposes a multi-stage refinement framework to generate fine-grained details by attending to relevant words in the

description. Recent GAN-based approaches [13, 10] propose object-driven, hierarchical approaches that explicitly model object instances within an image. [2] trains an ensemble of text-to-image diffusion models specialized for different synthesis stages. [15] explores diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance.

3 Model Architecture and Datasets

In this section, we talk about the architecture of the stable diffusion model [19] which we use for all our experiments. We further discuss the two datasets which we use to conduct our experiments.

3.1 Model

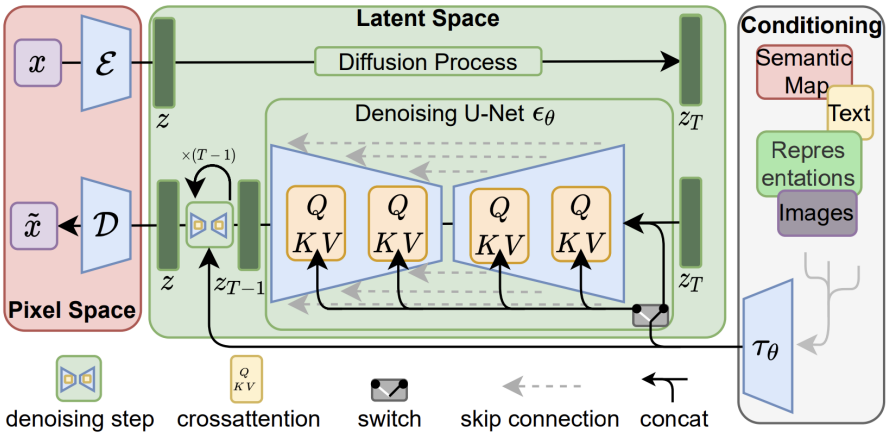

Fig. 2. Stable Diffusion

A variation of the diffusion model (DM) known as the latent diffusion model (LDM) is used by stable diffusion. Diffusion models, which were first used in 2015, are trained with the goal of eradicating repeated applications of Gaussian noise on training images, which can be compared to a series of denoising autoencoders. The variational autoencoder (VAE), U-Net, and an optional text encoder make up stable diffusion. The VAE encoder condenses the image from pixel space to a more basic semantic meaning in a less dimensional latent space. Forward diffusion involves repeatedly applying Gaussian noise to the compressed latent representation. The output of forward diffusion is denoised backwards by the U-Net block, which is built from a ResNet backbone, to produce latent representation. The VAE decoder then transforms the representation back into pixel

space to create the finished image. Flexible conditions can be applied to the denoising step based on a list of text, a picture, and other modalities. Through a cross-attention method, the encoded conditioning data is made available to denoising U-Nets. The fixed, pretrained CLIP ViT-L/14 text encoder converts text prompts to an embedding space for text conditioning. As a benefit of LDMs, researchers highlight to greater computing efficiency for training and generation. Fig 2 shows the visual depiction of conditioned LDMs either via concatenation or by more general cross-attention mechanism [19].

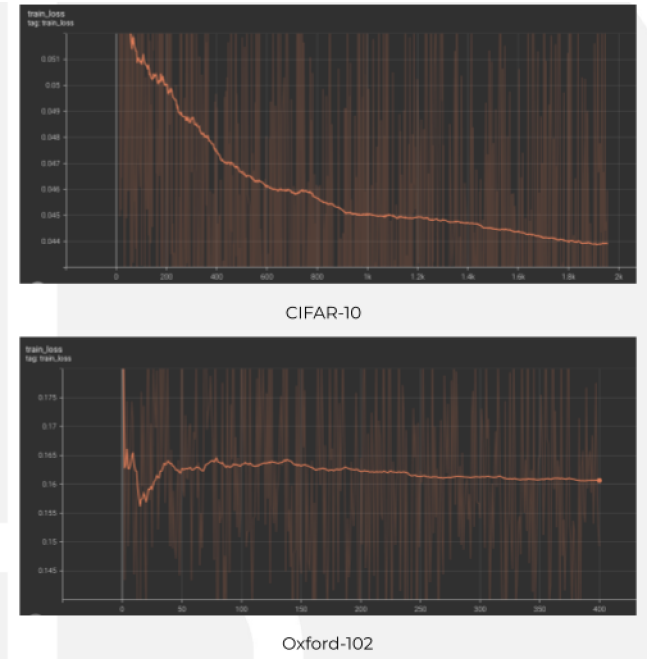

Fig. 3. Training Loss Vs Epoch Graph

3.2 Datasets

CIFAR-10 The CIFAR-10 dataset [11] consists of 6000 images per class in 10 classes totaling 60000 32x32 color images. The test-train split is taken as 1:5 and the class level balance is maintained across these splits.

Oxford-102 The Oxford Flowers 102 [16] collection consists of 102 categories of flowers that are frequently found in the United Kingdom. There are between 40 and 258 photos in each class. The photographs feature different poses, lighting, and scale. There are also categories with a lot of variety within the category and

Table 1. Stable Diffusion Hyper-Parameters

Hyper-Parameters	Values
Epochs	5
Learning Rate	1e-05
Batch Size	16
Mixed Precision	bf16
Resolution	512
LR Scheduler	Constant
Max Grad Norm	1
Gradient Accumulation Steps	8
Loss	Mean Squared Error

a lot of categories that are extremely similar. A training set, a validation set, and a test set are created from the dataset. Both the training set and validation set has ten photos for each class (totalling 1020 images each). The last 6149 pictures make up the test set (minimum 20 per class).

4 Experiments & Results

As discussed, we use the latest stable diffusion [1] model for fine-tuning on the above-mentioned datasets. The hyper-parameters used for fine-tuning can be found in Table 1. We do the fine-tuning only on the training set of both the datasets while generate images corresponding to the test set captions or text descriptions.

We generate both qualitative and quantitative performance results for our experiments. The qualitative results for CIFAR-10 dataset can be seen in Fig 4. The qualitative results for Oxford-102 dataset can be seen in Fig 5. The text below individual image is used as the input for our fine-tuned model. For the quantitative performance metric, we use FID [22] for both the datasets. For CIFAR-10 dataset, we also use the GAN Test [23] metric. As a part of the GAN Test, we basically use a model trained on the CIFAR-10 training dataset and evaluate it on the images generated corresponding to the textual inputs from test sets. We use pre-trained compact transformer model [8] for conducting the GAN Test. The reason why we don't conduct GAN Test for Oxford-102 is because the captions or natural language associated with each image does not provide enough context to form a correlation with semantics of label for each image. This makes the downstream classification task very difficult and the performance results obtained are not representative of the quality of images generated by the fine-tuned stable diffusion model. As we can see from both the qualitative and quantitative results, stable diffusion produces high quality images with respect to the dataset it is being fine-tuned.

[1] https://github.com/CompVis/stable-diffusion

Fig. 4. Qualitative Results for CIFAR-10

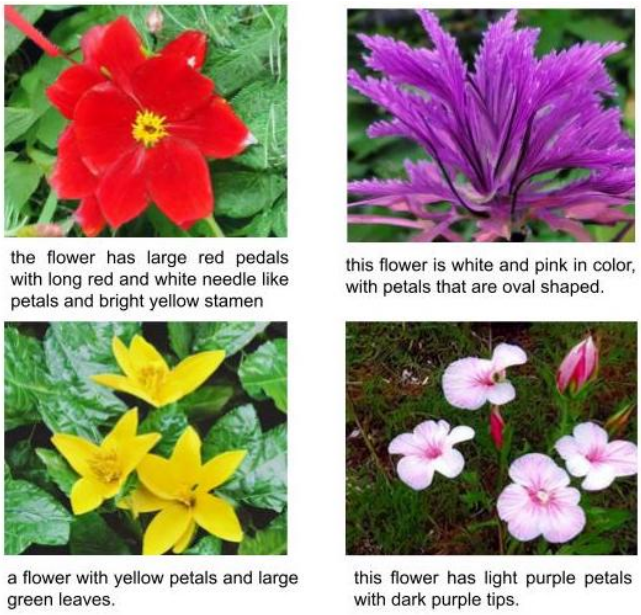

Fig. 5. Qualitative Results for Oxford-102

Table 2. Quantitative Evaluation Results

Dataset	Accuracy	FID
CIFAR-10	0.99	29.34
Oxford-102	-	28.21

5 Conclusion

This paper goes through the recent development of diffusion models in the literature and specifically investigates the use of the state-of-the-art stable diffusion model for text-to-image generation. Based on their mapped textual inputs, we fine-tune the model and generate high quality images for both CIFAR-10 and Oxford-102 datasets. We evaluate the quality of generated images using the traditional FID for both the datasets and GAN Test downstream classification for CIFAR-10.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Cvae-gan: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE international conference on computer vision. pp. 2745–2754 (2017)
4. Chen, L., Srivastava, S., Duan, Z., Xu, C.: Deep cross-modal audio-visual generation. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017. pp. 349–357 (2017)
5. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5630–5639 (2017)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
7. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
8. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H.: Escaping the big data paradigm with compact transformers. arXiv preprint arXiv:2104.05704 (2021)
9. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. J. Mach. Learn. Res. **23**, 47–1 (2022)

10. Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Text-to-image generation grounded by fine-grained user attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 237–246 (2021)
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
12. Li, B., Qi, X., Lukasiewicz, T., Torr, P.: Controllable text-to-image generation. Advances in Neural Information Processing Systems **32** (2019)
13. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12174–12182 (2019)
14. Liao, W., Hu, K., Yang, M.Y., Rosenhahn, B.: Text to image generation with semantic-spatial aware gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18187–18196 (2022)
15. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
16. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
17. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 49–58 (2016)
18. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Advances in neural information processing systems **29** (2016)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
20. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
21. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
22. Seitzer, M.: pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid (August 2020), version 0.2.1
23. Shmelkov, K., Schmid, C., Alahari, K.: How good is my gan? In: Proceedings of the European conference on computer vision (ECCV). pp. 213–229 (2018)
24. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200 (2016)
25. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
26. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1316–1324 (2018)
27. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: European conference on computer vision. pp. 776–791. Springer (2016)

28. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
29. Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., Mori, G.: Lifelong gan: Continual learning for conditional image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2759–2768 (2019)
30. Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., Guo, B.: Styleswin: Transformer-based gan for high-resolution image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11304–11314 (2022)
31. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 5907–5915 (2017)
32. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence **41**(8), 1947–1962 (2018)
33. Zhang, H., Dana, K.: Multi-style generative network for real-time transfer. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
34. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7151–7160 (2018)
35. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. Advances in neural information processing systems **30** (2017)
36. Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A.: A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1004–1013 (2018)