Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Text-to-Classic: A Diffusion Method for Classical Art Generation Based on Text

Yi Li

School of Science and Engineering, The Chinese University of Hong Kong Shenzhen, Shenzhen, 518172, China
yili24@link.cuhk.edu.cn

Abstract: Text-to-Image generation has recently become a hot research topic and diffusion models have achieved remarkable performance in this task. However, most previous researches aim at real scene generation. Few researches focus on classical art paintings. Besides, diffusion models are commonly heavy-weighted with a large number of parameters, which has a high computational cost. In this paper, we aim to solve the classical art paintings synthesis subtask. We propose a lightweight diffusion model Text-to-Classic(T2C) to synthesize classical art paintings according to text descriptions. Experiment results show that our method can achieve good performance with fewer parameters.

Keywords: Denoising Diffusion Probabilistic Model; Text-to-Image Generation; Art.

1. Introduction

With the progress in Artificial Intelligence (AI) and deep learning, researches on image synthesis have come into prominence. AI image synthesis is a task where AI learns to understand and reproduce images generated by humans. Many methods have achieved remarkable performances and generated images indistinguishable from the real by most people. While traditionally realistic images are synthesized from noise inputs, recently multimodal learning has become a hot spot and text-to-image synthesis is at the forefront. In text-image multimodal learning, AI learns to understand text-image correspondence. It combines natural language processing (NLP) and computer vision (CV). In the task of text-to-image synthesis, given a description in natural language, a realistic image matching the description will be generated.

A typical approach for text-to-image tasks is to use an NLP model as the encoder and an image synthesis model as the decoder. By such an encoder-decoder design, AI learns to combine both linguistic and visual information. While the correspondence between common-scene images and their descriptions is relatively simple, understanding artwork is much more challenging. Text-to-art, as a subtask of text-to-image synthesis, aims at training AI to understand art language. The goal of this task is to generate high-quality artworks from descriptions. Given its great value in social media, understanding artworks based on heuristics has been widely researched. It is easy for normal people to imagine common scenes based on descriptions since these elements are familiar in our daily life. However, envisaging art paintings is much harder for the public due to the lack of art knowledge. Understanding art seems to become the privilege of professional art critics.

To address such a problem, AI art understanding has become necessary. Neural style transfer (NST) aims to train AI to learn from style images. In NST, features from a "style" image and a 'content" image are extracted, and a stylized image is created by combining these features. In NST, AI learns to understand artworks from images with different styles. However, it is hard for the public to find the style images that exactly meet their requirements. Compared with style images, text descriptions are more obtainable heuristics. Text-to-art synthesis, where AI understands art by natural language, can convert abstract art words into vivid art paintings. Most multimodal works do not treat text-to-art as an independent task. For those works involving art synthesis, it only includes art figures as part of the training data, and mainly focuses on fine art.

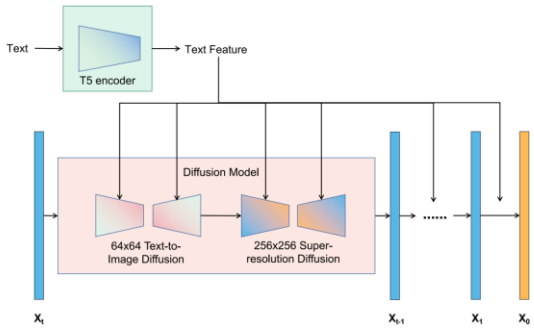

Figure 1. Overall pipeline of our proposed method

We use Imagen as our framwork. A text description is first encoded to high-level features by T5 encoder. The text features are then fed to the two-diffusion model to guide the generation.

In this paper, we propose text-to-classic, which focuses on multimodal classical art painting synthesis. We use Imagen [1] as our pipeline, and compress the network into a lightweight structure. Experiments show that the compressed pipeline can still achieve a reasonable synthesis quality, but with a much lower number of parameters and computation cost.)

In summary, our contributions are:

We propose to apply diffusion models in text-to-classic image synthesis task, which is rarely discussed by previous research.

We compress the network of Imagen and reduce computational costs. The light-weighted network is proved to be comparable with the original heavy-weight pipeline by experiments.

2. Related works

Since few researches focus on the classic art paintings synthesis task, in this section, the more general task of Text-

to-Image generation will be discussed. In this task, given a text description, an image with high fidelity should be generated according to the description. It is a combination of CV and NLP and has drawn much of the researchers' attention.

2.1. GAN-based generation

GAN-INT-CLS [2] is the first work adopting a conditional Generative Adversarial Network for text-to-image generation. It uses a deep convolutional-recurrent text encoder to encode the input text into a 128-dimensional vector. The text vector is then concatenated with the noisy latent code as the input to the generator. [3] combines both text descriptions and classification labels to guide the generation. [4] proposes a cascaded two-stage GAN to first generate an image with low resolution and then conduct super-resolution. [5] leverages a mapping network and a ternary mutual information loss to combine text and latent information and improve interpretability.

2.2. Diffusion-based generation

Ever since proposed by [6], image generation models based on diffusion have achieved remarkable performance. Similar to GAN-based text-to-image synthesis methods, diffusion-based methods use a natural language processing model to encode text information, and the difference lies in the decoder. Instead of using GAN, diffusion-based methods exploit the Denoising Diffusion Probabilistic method (DDPM) to sample images. Following [6], most diffusion models are based on the Unet structure. [7] provides classification labels to the diffusion model, and proves that diffusion models outperform traditional GAN models by experiments. [8] adopts a cascaded structure with a low-resolution base diffusion generator followed by several super-resolution diffusion models. Similar to [6], the work of [8] utilizes classification labels as conditions to the network. [9] proposes to use texts for conditioning, aiming at the text-to-image generation task. [1] achieves the state-of-the-art performance in Text-to-Image synthesis. Therefore, in this paper, we use Imagen proposed by [1] as our framework to solve the text-to-classic problem. However, the origin Imagen is a heavy-weighted network, with high computational costs in training and sampling. Therefore, we compress the network, and experiment results show that the light-weighted network has a comparable performance with the original one.

3. Methodology

To solve the classical image generation problem, we propose Text-to-Classic(T2C), a lightweight Text-to-Image generator based on the diffusion model. We leverage Imagen [1] as the framework for Text-to-Classic image synthesis. Imagen is a Text-to-Image diffusion model consisting of a frozen generic large language model as the encoder and a cascaded diffusion model as the decoder.

3.1. Pretrained text encoder

Text encoder can be trained on pure text data or image-text pairs. While NLP models (e.g. CLIP) trained on text-image-pair can extract vision-language correspondence, which is beneficial to Text-to-Image synthesis, image-text datasets are always much smaller compared with pure text datasets. Alex et al [9]. compares different language encoders and found that models trained on large pure text datasets outperform those trained on text-image pairs in terms of sample quality. Specifically, T5 model is experimentally proven to be the most efficient text encoder. Text-to-Text Transfer Transformer(T5) is a transformer model following the classical encoder-decoder structure proposed by [10] and is trained on their Colossal Clean Crawled Corpus(C4) dataset. In the rest of this section, the encoder-decoder transformer structure will be briefly introduced.

a) Self-Attention Mechanism: Given an input sequence X = x1, x2, ..., xt, a Self-attention function first maps each element xi of the sequence into three feature vectors: query Qi, key Ki and value Vi.

$$Q_i = W_Q x_i \quad (1)$$

$$K_i = W_K x_i \quad (2)$$

$$V_i = W_V x_i \quad (3)$$

The output sequence Y = {y1, y2, ..., yt} of the Self-Attention function has the same length as the input. Each element in the output sequence is computed as the weighted sum of all values Vi, and the weights are calculated by quering all keys:

$$y_i = \sum_{j=1}^{t} \alpha_{ij} v_j \quad (4)$$

where weights αij can be computed by:

$$\alpha_{ij} = \text{Softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k=1}^{t} \exp(e_{ik})} \quad (5)$$

$$e_{ij} = \frac{Q_i^T K_j}{\sqrt{d_k}} \quad (6)$$

where dk is the number of dimension of keys.

b) Multi-Head Attention subblock: Multi-Head attention subblock comprises multiple Self-Attention functions. The final output is the weighted sum of outputs from n Self-Attention funtions:

$$Yi = \text{Attention}(X) \quad (7)$$

$$Y = \sum_{i=1}^{n} Yi \quad (8)$$

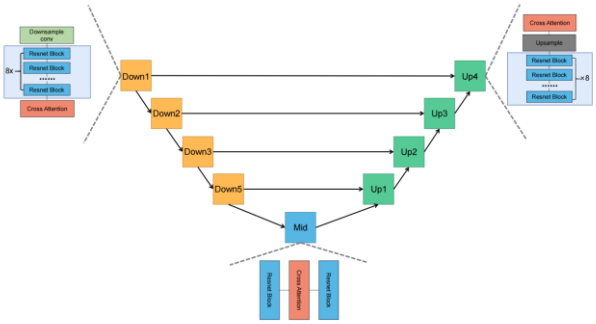

Figure 2. Structure of the 4-layer Unets

Each downsample layer consists of a convolution layer for pre-downsampling, 8 Resnet blocks, and a cross-attention layer. The middle layer is composed of 2 Resnet blocks and a cross-attention layer. Each upsample layer comprises 8 Resnet blocks, a upsample block, and a cross-attention block. The skip connecting is between each Resnet block of downsample layers and up sample layers. Text conditions are added to the network in cross-attention block.

c) Encoder: The encoder is a stack of identical blocks. Each basic block of the transformer mainly comprises two subblocks: a Multi-Head Attention block and a Feed Forward block. The Feed Forward block is simply a fully connected layer. Both subblocks are followed by a residual skip connection and a layer normalization. The block of T5 is similar to the block proposed in [10], with the modification in the skip connection and normalization. In T5 model, the layer norm is simplified by only keeping the rescaled without the additive bias. The residual skip connection is after the

normalization, so the output of each subblock is x+SimpLayerNorm (subblock(x)). Besides. dropout is applied within

the Feed Forward network.

The input text tokens are first mapped to sequence embedding and added with positional embedding with the same shape, which is then passed to the encoder. After several transformer blocks, high-level features F = {Qe, Ke, Ve} are extracted.

d) Decoder: The decoder also comprises several identical blocks. The structures of these blocks are similar to those in the encoder, but with an addition Multi-Head Attention subblock appended before the Feed Forward subblock. While the first Multi-Head Attention function takes the output embedding as the input, the addition attention function put attention to both features from the last decoder block as well as the high-level features from the encoder. Specifically, it attends on the query from the first attention function and the key-value pair from the high-level encoder features {Qd, Ke, Ve}. The final output of the decoder is passed to a softmax layer, indicating the probability of each word token.

Imagen exploits the encoder of T5 transformer to convert text tokens to high-level features.

3.2. Diffusion-Based Decoder

1) Denoising Diffusion Probabilistic Models: DDPM comprises a forward noising process and a reverse denoising process. Given a data x0 in distribution q (x0): x0 ~ q (x0), the forward process gradually adds gaussian noise to the data, producing a sequence from x1 to xT:

$$q(x_{1:T}|x_0) = \prod_1^T q(x_t|x_{t-1}) \quad (9)$$

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (10)$$

The sequence is a Markov chain and it is proved that if T is sufficiently large, xT is an isotropic Gaussian distribution.

Therefore, if given the reverse distribution p(xt-1|xt) for every time step t in the sequence, the original data x0 can be recovered from a pure Gaussian noise through a sequence of sampling. While the reverse distribution p(xt-1|xt) depends on the forward process, a neural network can be used to infer it without forwarding information:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)) \quad (11)$$

where μθ, Σθ are the parameters predicted by the network.

In training, the goal is to adjust parameter θ to make the predicted distribution of reversed Markov chain pθ(x0, x1, ..., xT) close to the groundtruth distribution q(x0, x1, ..., xT). This can be achieved by minimizing the variational lower bound (VLB):

$$L_{vlb} = \sum_{i=0}^T L_i \quad (12)$$

where

$$L_0 = -\log p_\theta(x_0|x_1) \quad (13)$$

$$L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t) \quad (14)$$

$$L_T = D_{KL}(q(x_T|x_0)||p(x_T)) \quad (15)$$

However, in practice, a simplified training objective is used. Instead of predicting the mean and variance of the distribution in each step, the network could also predict the total noise ∈ added to the original data at time step i. The objective becomes:

$$L_{simple}(\theta) = E[||\epsilon - \epsilon_\theta(x_t, t)||^2] \quad (16)$$

where ∈ θ is the predicted noise, and ∈ is the groundtruth noise randomly added to the origin data x0 to generate the data xi in step i during training:

$$x_i = \sqrt{\alpha_i}x_0 + \frac{\beta_i}{\sqrt{1-\alpha_i}}\epsilon \quad (17)$$

$$\alpha_i = 1 - \beta_i \quad (18)$$

$$\bar{\alpha}_i = \prod_{j=0}^i \alpha_j \quad (19)$$

A better sample quality can be obtained by minimizing such a simplified objective function, according to Ho et al. [11].

Similar to GAN-based image synthesis, diffusion-based image generation can be conditional. Conditional diffusion models predict noise ∈ θ not only based on the sampled sequence but also under the guidance of conditioning information c. Therefore, the optimization objective becomes:

$$L_{simple}(\theta) = E[||\epsilon - \epsilon_\theta(x_t, t, c)||^2] \quad (20)$$

2) Efficient Unet Decoder: The decoder in Imagen conditions the text embedding extracted by the encoder and synthesizes an image matching the description. It follows a cascaded structure consisting of three diffusion models. The base model maps the text embedding into a low resolution 64 x 64 image, which is then passed to a 64×64→256×256 and a 256×256→1024×1024 diffusion super-resolution model. All three models are based on an Unet architecture. The base Unet takes the noisy data xt in a sampling sequence as an input, and the output is the predicted noise ∈ θ. For the two super-resolution Unets, noisy data xt is first concatenated with the resized low-resolution image and then fed into the networks.

The cross-attention mechanism is added to all three Unets to condition text embeddings. In attention heads, queries are derived from features extracted from xt while text embedding provides key-value pairs. We only utilize the first two Unets of Imagen, so the size of generated paintings is 256 × 256.

In the work of [1], Unets are heavy-weighted with a large number of parameters: the 64 x 64 base Unet contains 2B parameters while the 64 × 64 → 256 × 256 super-resolution Unet contains 400M parameters. Such large networks not only result in difficulties in training and sampling but also cause overfitting problem when training on small datasets. For Text-to-Classic synthesis, image-description pairs are much less than common scene datasets such as COCO [11]. Therefore, a lighter-weight Imagen is needed for this task. We compress the network by dimensionality shrinking of feature maps in the Unet pyramid. We reduce the dimension of feature maps in two ways: decreasing the base dimension, and decrementing the multiplication factor in each layer. By shrinking the dimension of feature maps, the network can be efficiently compressed while the multiscale information can be preserved.

For columns, from left to right are: text description of art paintings, ground truth image of paintings, 64 × 64 intermediary image output by the base Unet, and the final 256 × 256 synthesized images.

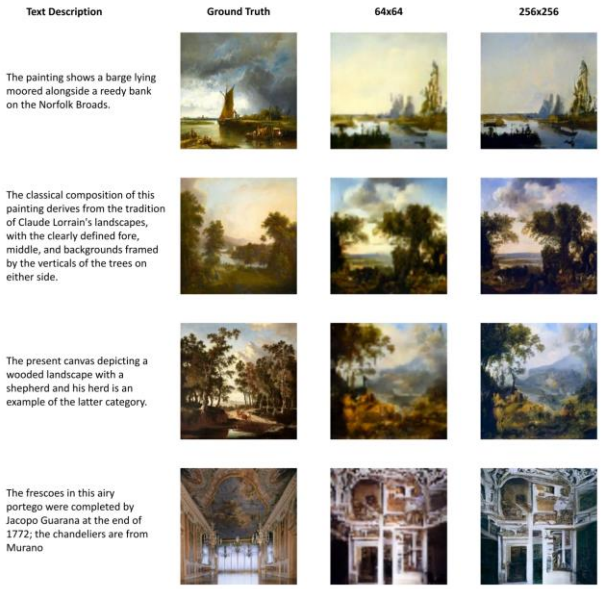

Figure 3. Qualitative results of our method

4. Experiments

4.1. Dataset

Our T2C model is trained and tested in SemArt dataset [12]. SemArt is a collection of classic art painting images. It contains 21,384 samples and each sample consists of a classic art painting and its information in texts. The 21,384 samples are split into training, validation, and test sets according to a ratio of 20:1:1, following the official partition [12]. The number of samples in each set is summarized in table 1.

Table 1. Fid achieved by base Unet with different base dimension

Base dimension	Number of parameters	FID
32	8M	110.8821
64	28M	78.3743
128	106M	82.6170
256	410M	71.2723
320	635M	71.3565

To measure the performance of our T2C quantitatively, we adopt the popular metric Fr chet inception distance (FID). We use the test set to calculate FID.

4.2. Experiment Setup

Different dimensionality configurations of feature maps are tested, to measure the effect of network compression. The 64 × 64 base Unet and the 64 × 64 → 256 × 256 super-resolution Unet are tested separately. Pyramids of both Unets have 4 layers, and feature maps of the first layer have a base dimension. For each of the rest of the layers, the dimension of feature maps is the multiple of those in its last layer.

$$d_i = \alpha_i \times d_{i-1},\ i = 2, 3, 4 \qquad (21)$$

where di is the dimension of feature maps in layer-i, and αi is the multiplication factor. The multiplication factor is set to be α2 = 2, α3 = 3, α4 = 4. Different base dimensions d1 are explored. Two Unets are evaluated separately, and the super-resolution Unet samples on resized 64×64 ground-truth images.

Table 2. Fid achieved by super-resolution Unet with different base dimension

Base dimension	Number of parameters	FID
32	16M	110.8127
64	61M	68.2957
128	241M	64.0859
192	538M	62.6451

4.3. Quantitative Results

As is shown in Table 1, for the base Text-to-Image Unet, shrinking the base dimension leads to a small loss in synthesis quality. However, the FID value is still reasonable when the base dimension is no less than 64.

The performance of Unet 2 has a similar tendency to those of Unet 1, according to the results in Table 2. Compressing the network only results in a slight drop in FID when the base dimension is greater than 32. Trading off synthesis quality and the number of parameters, the combination of a 64-dimensional base Unet and a 64-dimensional super-resolution Unet is the best choice.

4.4. Qualitative Results

In this section, qualitative results of the 64-64 combination are demonstrated. Such a combination has only 89M parameters, while the heaviest 320-192 combination consists of 1173M parameters. The 64-64 combination is only 7.5% of the heaviest combination. As is shown in figure 3, despite the relatively small number of parameters, our model can still extract information from the challenging text description and synthesize a reasonable art painting according to the description.

5. Conclusion

Previous researches on Text-to-Image generation pays little attention to classical art synthesis, this paper fills such a gap.

In this paper, we propose a light-weighted diffusion model T2C for Text-to-Classic generation. We adopt Imagen as our framework. Given a text description, it is first encoded into high-level features by T5 transformer encoder. These features then serve as conditions to guide the cascaded diffusion-based decoder. We compress the network for more efficient training and sampling. Different experiments are conducted to improve the model and reduce the computational cost. Experiment results show that our lightweight model can still achieve a reasonable performance.

References

[1] Saharia, C., Chan, W., Saxena, S., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Arxiv Preprint, 2205.11487.

[2] Reed, S., Akata, Z., Yan, X., et al. (2016). Generative adversarial text to image synthesis. PMLR, 1060-1069.

[3] Dash, A., Gamboa, J. C. B., Ahmed, S., et al. (2017). Tac-gan-text conditioned auxiliary classifier generative adversarial network. Arxiv Preprint, 1703.06412.

[4] Zhang, H., Xu, T., Li, H., et al. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. Proceedings of the IEEE international conference on computer vision, 5907-5915.

[5] Yuan, M., Peng, Y. (2019). Bridge-GAN: Interpretable representation learning for text-to-image synthesis. IEEE Transactions on Circuits and Systems for Video Technology, 30(11):4258-4268.

[6] Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840-6851.

[7] Dhariwal, P., Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34:8780-8794.

[8] Ho, J., Saharia, C., Chan, W., et al. (2022). Cascaded Diffusion Models for High Fidelity Image Generation. Journal of Machine Learning Research, 23(47):1-33.

[9] Nichol, A., Dhariwal, P., Ramesh, A., et al. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. Arxiv Preprint, 2112.10741.

[10] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[11] Lin, T., Maire, M., Belongie, S., et al. (2014). Microsoft coco: Common objects in context. In : Computer Vision–ECCV 2014: 13th European Conference. Zurich, Switzerland. 13:740-755.

[12] Garcia, N., Vogiatzis, G. (2018). How to read paintings: semantic art understanding with multi-modal retrieval. Proceedings of the European Conference on Computer Vision (ECCV) Workshops , 0-0.r